Universiteit Leiden

Measuring ESG Content in Financial Text with Frozen LLM Activations: Concept Vectors and Linear Probing

Luc Hazenoot (s3712206)

Supervisors:
Dr. A.H. Zohrehvand & Dr. Z. Ren

Bachelor Thesis– Informatica en Economie

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

February 2026

Abstract

Financial texts contain valuable data relevant to investors and analysts in assessing concepts such as firm performance, risk and sustainability; however, measuring how much of a text is about a certain concept remains an open challenge. Recent work has shown that a gap exists in what Large Language Models (LLMs) know internally versus what they express in their response. This thesis utilises this fact to build a cheap, off-the-shelf measure of Environmental, Social and Governance (ESG) content in financial text by monitoring the internal activations of frozen, out-of-the-box LLMs. Using a human-annotated ESG dataset, we extract concept vectors via the Recursive Feature Machine (RFM) algorithm and via linear probing, and compare these against an embedding method baseline. We find that these methods come close to fully fine-tuned models without any task-specific fine-tuning. Among the activation-based methods, linear probing performed best, while the RFM algorithm additionally provides a continuous score intended to reflect how strongly a concept is present in a text, rather than only whether it is present.

1 Introduction

Financial texts contain valuable data relevant to investors and analysts in assessing concepts such as firm performance, risk and sustainability. Automating these analyses via Natural Language Processing (NLP) has been a transformative development in finance (Gentzkow et al., 2019, Loughran and McDonald, 2011). However, measuring these concepts remains an open challenge. Recent work has shown that a gap exists in what Large Language Models (LLMs) know internally versus what they express in their responses (Beaglehole, Radhakrishnan, et al., 2026, Burns et al., 2023, Azaria and Mitchell, 2023, Marks and Tegmark, 2024). Nevertheless, a comprehensive survey conducted by Du et al., 2025 covers ten major applications of NLP in finance, none of which involve monitoring the internal activations of the LLMs.

Significant work has gone into optimising the performance of an LLM. Output-focused methods guide the model into stepwise reasoning leading to the discovery of "chain-of-thought" prompting (Wei et al., 2022). While other methods focus on the internal representation of LLMs. The linear representation hypothesis proposes that although a model consists of complex non-linear computations, concepts are linearly accessible as directions in activation space (Park et al., 2024, Marks and Tegmark, 2024). Beaglehole, Radhakrishnan, et al., 2026 further strengthens this hypothesis by finding what they call concept vectors.

With current methods, it is difficult to measure how much of a text is about a certain topic, rather than only whether the topic is present. This matters because knowing how much of a text concerns a topic tells us how much, and in what area, a sentiment is valuable.

This thesis aims to find out whether the fact that an LLM knows more internally than it expresses in its output can be utilised to better analyse pieces of financial text. We do this by testing linear probing and the RFM concept-vector method against our baselines. Our expectation is that the findings of Beaglehole, Radhakrishnan, et al., 2026 hold, and that their method of using the RFM algorithm to find concept vectors applies outside their researched domain onto ours, making RFM the strongest of our activation-based methods for analysing financial text. Furthermore, we expect that projecting the concept vectors onto new activations provides a continuous score representing how much of a concept is present in new sentences. In doing so, we contribute a

cheap, off-the-shelf measure of ESG content that operates on frozen, out-of-the-box LLMs and comes close to fully fine-tuned models without any task-specific fine-tuning.

2 Theoretical Background

2.1 Model Architecture

Modern decoder-only LLMs utilise a transformer architecture introduced by Vaswani et al., 2017. Transformer models are built out of a stack of identical computational blocks, also referred to as the hidden layers. Each block consists of a self-attention layer followed by a feed-forward layer. In decoder-only models, the self-attention layers are causal: when the representation of a token is updated, it can only see itself and the tokens that come before it, never the ones that have yet to follow. The output of each block is a learned representation that captures the information the model has built up so far, which is then passed as input to the next block. The number of blocks inside decoder-only LLMs is model-specific, with some containing 30 blocks or fewer while others consist of 80 or more.

2.1.1 Internal activations

When the model processes a prompt, the input text is first tokenised, splitting it into a sequence of tokens that are converted into numerical vector representations (Vaswani et al., 2017). Each token vector is then passed through the model's blocks. After being processed by a block's self-attention and feed-forward layers, the result is a new learned representation of the token, referred to as the hidden state or activation. These activations are passed as input to the next block, and this iterative process is called inference.

2.1.2 Activation space

The activations are in a high-dimensional space, typically of several thousand dimensions, called the activation space. These activations represent the internal representation of the model's knowledge up until that point (Belinkov, 2022). Each block has its own activation space and encodes different kinds of information. Early blocks tend to capture low-level features, while middle to late layers capture more abstract ones (Tenney et al., 2019).

2.2 Concept Vector

A concept vector is a learned direction in the activation space of a single layer of a model. When correctly found, it should point towards the direction in the activation space that captures the most important features for predicting the concept. A concept vector can be used in two ways: to steer a model toward the concept by adding the vector to new activations, or to monitor how active the concept is within new activations.

2.2.1 Monitoring

Once a concept vector has been identified, an LLM can be monitored by looking at how active the concept is within the model's internal processing. This is beneficial because research has shown

that a model knows more than it expresses in its response (Beaglehole, Radhakrishnan, et al., 2026, Burns et al., 2023, Azaria and Mitchell, 2023), so by looking at the internal activations, it is possible to interpret what a model knows. This has been demonstrated to be a cost-efficient way of verifying the truthfulness of a model, previous methods relied on separate models to check for hallucination and accuracy. Yet, monitoring can run on a single GPU within a minute. (Beaglehole, Radhakrishnan, et al., 2026, Zou et al., 2023)

2.3 Recursive Feature Machine

The Recursive Feature Machine (RFM) algorithm is a supervised model developed by Beaglehole, Holzmüller, et al., 2026. It seeks to find a separation between classes by using non-linearity on the labelled activations of LLMs as input. It does so by computing the Average Gradient Outer Product (AGOP) matrix, which weighs which features are most important for telling the classes apart. This matrix is recalculated at the end of each iteration, and as the algorithm repeats, it keeps updating the matrix to capture which features matter most. Beaglehole, Radhakrishnan, et al., 2026 found that this same AGOP matrix can also be used to extract concept vectors. By applying eigendecomposition on the AGOP matrix, you find the vectors that best distinguish between the classes. Using eigendecomposition, the top n eigenvectors are extracted from the matrix. These vectors are what become the concept vectors. Finding the concept vectors this way was not the original purpose of the RFM algorithm; it is a by-product of how the AGOP matrix captures the most important directions for separating the classes.

2.4 Linear Probing

Linear probing is a supervised method that uses lightweight linear classifiers to interpret the activation space of a frozen LLM for classification tasks (Alain and Bengio, 2017, Belinkov, 2022). The linear classifier is trained on labelled activations to find the direction that best separates the desired concept from its absence. The classifier's prediction is then interpreted as how strongly the concept is represented in that layer (Belinkov, 2022, Hewitt and Liang, 2019). Common models chosen include linear ridge regression (Hoerl and Kennard, 1970) and logistic regression.

2.5 Token Pooling

After a prompt of N tokens is processed by a block inside the model, the block outputs a sequence of N activation vectors, one per token. Methods like RFM and Linear probing usually require a single fixed-size vector per layer rather than the entire sequence of N activation vectors. Token pooling refers to the method used to transform the sequence of N activation vectors into one. Three common strategies are last-token activation, mean pooling, and max pooling (Reimers and Gurevych, 2019).

Last-token activation uses the activation vector of the final token in the sequence as the prompt's representation. In decoder-only LLMs, tokens can see the values of the tokens that precede them. The last token has therefore seen the entire sequence and can thus be interpreted as containing a summary of the entire sequence.

Mean pooling takes the mean of the entire sequence of N activation vectors, producing a single vector. It treats every token equally, which allows signals to remain visible but dilutes signals that

are only active in a few tokens.

Max pooling takes the maximum value per signal out of the sequence of n token activation vectors, producing a single vector. This creates an emphasis on the stronger activation along each direction, which strengthens signals that are only active in a few tokens, but this also results in it being more sensitive to outliers than mean pooling.

3 Methods

In this section, we will describe the experimental methods used to evaluate whether internal activation can be used as an improvement on current embedding models.

As a baseline, we use an embedding model. Embedding models are a standard way to turn text into dense feature vectors that can be used for classification (Reimers and Gurevych, 2019). This allows us to get a realistic view of how conventional NLP methods perform on our task, which we can then compare against our activation-based methods.

For the embedding baseline, we use the Qwen-3-embedding-8b model. We choose this model because it is a recent embedding model that is based on the Qwen-3 family that we also use for our activation-based experiments. We take the embeddings and use them as input features in a logistic regression model in order to predict the binary labels. We do the prediction by taking the mean of the metrics over a 5-fold cross-validation, which reduces potential split bias that the dataset might introduce, and keeps the evaluation consistent with our other experiments.

For our experiments, we used a range of decoder-only models to test how different families and sizes perform.

Llama-3.1-8b-it is a decoder-only model consisting of 32 layers containing a total of 8 billion parameters. Previous research has already shown that this model performs well on activation-based tasks (Beaglehole, Radhakrishnan, et al., 2026).

From the Qwen-3 family we use two models, Qwen-3-8b-it and Qwen-3-14b-it. Both models are decoder-only models, the 8b model consisting of 36 layers containing a total of 8 billion parameters and the 14b model consisting of 40 layers and 14 billion parameters. These models were released in 2025 and have been demonstrated to perform well on reasoning tasks. Our expectation is that these models should be better at understanding more complex questions.

Gemma-4-31b-it is a decoder-only model consisting of 60 layers containing a total of 31 billion parameters. Since this is a larger model we will load it in 8-bit quantization. This model was released in 2026 and is the biggest newest model we use. Our expectations are that this will be the best model for complex tasks and concept vectors.

3.1 Dataset

The dataset we use is a human-annotated ESG dataset by Schimanski et al., 2024, containing 2000 sentences per ESG pillar: Environmental, Social, Governance. The source is a corpus of 13.8 million sentences extracted from: annual reports, responsibility reports, sustainability reports and news articles. Each dataset is built out of the following: 1000 shared sentences, 750 domain-specific sentences and 250 general language sentences. Each sentence contains a binary label 0 if the domain is not present and 1 if it is.

The 1000 sentences are shared between the three datasets, and each sentence is annotated three times for its respective pillar. A shared sentence, therefore, might contain multiple positive labels.

The 750 domain-specific sentences are subdivided into two categories: 500 sentences filtered by the pillar keywords and 250 with no keyword filter. Sentences filtered by the pillars keyword are not by default true positives, for example, a sentence containing the keyword "green" could be about a traffic light and not related to the environment.

The last 250 general sentences are unfiltered sentences from the corpus, mostly containing negative sentences that are unrelated to the specific pillar.

Three expert annotators then labelled each sentence following a shared rubric. Their inter-annotator agreement is shown in Table 1.

Domain	**Fleiss' kappa**	**3-of-3 agreement**
Environmental	0.906	89.8%
Social	0.926	92.3%
Governance	0.867	87.0%

Table 1: Inter-annotator agreement across the three ESG domains.

With the resulting label distribution per dataset shown in Table 2

Dataset	**Yes (True)**	**No (False)**	**Total**
Environmental	658	1,342	2,000
Social	804	1,196	2,000
Governance	538	1,462	2,000

Table 2: Approximate label distribution in the 2k expert-annotated ESG datasets (read from Figure 1).

3.2 Experiments

3.2.1 RFM

Our primary experiments use the RFM algorithm introduced by Beaglehole, Holzmüller, et al., 2026; Beaglehole, Radhakrishnan, et al., 2026. We first randomly shuffle the dataset so that the labels are evenly distributed, and then divide it into three splits: a training split of 1000 sentences, a validation split of 500 sentences, and a test split of 500 sentences. Every sentence is wrapped in a fixed statement whose purpose is to guide the model towards the concept we want to extract.

The next step is to extract the concept vectors. Using the training split, we extract the activations of each layer of the model and use them as input to the RFM algorithm. From the resulting AGOP matrix, we take the top $N = 5$ eigenvectors by eigendecomposition and use them as our concept vectors.

To score a new sentence, we compute the cosine similarity between its activations and the concept vectors. For each concept vector n, we first compute the per-layer cosine similarity between the activation and the concept vector, and then average these per-layer values into a single score

for that concept vector. This is repeated for all N concept vectors, and the N resulting scores are averaged into one final score per sentence. We compute this final score for every sentence in the split, and then scale the scores to fit between -1 and 1 for readability. An optimal threshold is then found on the scores in order to predict the label. From this we calculate the following metrics: Area Under the Curve (AUC), accuracy, F1, precision and recall. These metrics serve as a check of how good the found concept vectors are.

Because concept vectors differ based on what wrapper was used, we evaluate the wrappers against each other based on accuracy. This is because Schimanski et al., 2024 evaluated their baseline models on this metric, which gives us a direct comparison against them. We select the wrapper with the highest accuracy on the validation split, and then finally evaluate that single wrapper's concept vectors once on the held-out test split to obtain our final score and metrics.

We report the metrics of the first concept vector as well as the mean of the top 5 concept vectors. The concept vectors are the top eigenvectors of the AGOP matrix, ordered by eigenvalue, so the first concept vector is the one that separates the classes the strongest. We report it on its own, alongside the mean of the top 5, in order to see how much of the performance comes from this single strongest direction compared to using multiple concept vectors together, which may capture more of a concept when it is spread across several directions.

We also show the continuous scaled scores and their distribution. These scores reflect the similarity between the concept vectors and new sentences, where a higher score corresponds to a stronger alignment with the concept. This score is intended to reflect how much the concept is present in a new sentence, rather than only whether it is present.

3.2.2 Linear Probing

For the linear probing approach, we use a stacked generalisation architecture. In the first stage, a separate RidgeClassifier with strong L2 regularisation is trained on each selected layer's activations. This is because of the high dimensionality of the activation space. This produces a per-layer decision score. In the second stage, a logistic regression meta-learner is trained on the aggregated per-layer scores to produce the final prediction. To avoid data leakage when training the logistic model, we use a nested cross-validation procedure. The outer loop is a 5-fold stratified cross-validation used to compute the final score. Within each outer training fold, an inner 5-fold stratified cross-validation generates out-of-fold predictions. This means that each training example receives a decision score from a probe that was not exposed to it during fitting. These out-of-fold scores serve as logistic regression model training features. Inner fold probes are then refit on the full outer training fold to produce decision scores on the outer test set. Standardisation is also performed inside each respective inner and outer fold so that scaling statistics are computed only on the relevant training slice in both the inner and outer loops to avoid minor data leakage. The fitted logistic regression model additionally provides a per-layer importance signal through its coefficients, which are retained for analysis.

Since each layer of an LLM produces one activation vector per input token, a pooling strategy is required to obtain a single fixed-size representation per example. We compare three strategies: last-token, mean pooling, and max pooling. Each strategy is evaluated independently through the full probing pipeline described above.

3.3 Comparison of methods

Both Linear Probing and RFM are supervised methods that aim to find a direction in the activation space that best separates prompts where the target concept is present from those where it is not. Linear probing trains a linear classifier on the activations; the direction normal to its decision boundary is the direction that best separates the two classes (Belinkov, 2022). RFM does this differently: it trains a non-linear predictor on the activations and extracts the top eigenvectors of the AGOP matrix, which are taken as the concept vectors. To detect the concept in a new prompt, these per-layer concept vectors are projected onto the prompt's activations at each layer, and the resulting similarities are aggregated into a single score that is thresholded to make the prediction. Both methods thus find a direction in the activation space, one linearly and the other non-linearly.

4 Results

Tables 3, 4 and 5 provide a full overview of the classification results for every method on the Environmental, Social and Governance pillars, using Area Under the Curve (AUC), accuracy (ACC), F1, precision (Pr) and recall (Re). Metrics for the embedding baseline and linear probing are the mean over a 5-fold cross-validation; RFM metrics are computed once on the held-out test split, using the wrapper selected on the validation split. Figure 1 shows the distribution of the continuous RFM scores.

4.1 Activation-based comparison

Across all datasets, linear probing achieved the highest results of the activation-based methods and comes close to the fine-tuned models from Schimanski et al., 2024.

On the Environmental dataset, the best linear probing experiment, which used last token activation on the Qwen-3-8b model, scored 0.9505 accuracy, coming within 0.6 percentage points of the fine-tuned EnvRoBERTa, which achieved 0.9565 on accuracy. The best RFM configuration, which used Gemma-4-31b and the first concept vector, scored 0.946 on accuracy. The baseline embedding method scored 0.9425.

On the Social dataset, the baseline embedding method scored 0.9245 on accuracy. The best linear probing method, using max pooling and Qwen-3-14b, almost evens that scoring 0.05 percentage points lower with an 0.9240 score on accuracy. The best RFM algorithm, which used the Gemma-4-31b and the first concept vector, scored 0.886. The best fine-tuned model SocRoBERTa scored 0.9341.

On the Governance dataset, the baseline embedding method scored 0.8355 on accuracy. Using linear probing, with max pooling and the Qwen-3-8b model, we achieved a score of 0.8755. The RFM method with Gemma-4-31b and first concept vector, scored 0.866. The best fine-tuned model GovDistilRoBERTa, scored 0.8966.

Across all pillars, the fine-tuned models from Schimanski et al., 2024 outperform all other methods. From the activation-based methods, the best linear probing method achieves higher metrics than the best RFM methods.

For linear probing there is no single best pooling strategy, last token pooling is best on Environmental, while max token pooling is best on Social and Governance. The same holds for model selection on linear probing as Qwen-3-8b-it is the best on Environmental and Governance, while Qwen-3-14b-it is the best on Social.

For RFM, the first concept vector matches or beats the mean of the top five concept vectors on accuracy in every pillar. There is also a pattern in model selection as Gemma-4-31b-it outperforms all other models on accuracy across every pillar.

Table 3: Classification Results on the Environmental Pillar of the ESG Dataset. Within each method, the best value per metric is shown in bold.

Category	Method	AUC	ACC	F1	Pr	Re
2*Baselines	Embedding method	0.9806 ± 0.0065	0.9425 ± 0.0047	0.9157 ± 0.0080	0.8841 ± 0.0066	0.9498 ± 0.0225
21*Experiment	*RFM*					
	Gemma-4-31b-it					
	First concept vector	**0.985**	**0.946**	**0.924**	0.912	0.938
	Mean of 5 concept vectors	0.984	0.944	0.923	0.894	**0.955**
	Llama-3.1-8B					
	First concept vector	0.962	0.902	0.860	0.867	0.852
	Mean of 5 concept vectors	0.931	0.872	0.818	0.818	0.818
	Qwen-3-8B-it					
	First concept vector	0.965	0.900	0.846	**0.926**	0.778
	Mean of 5 concept vectors	0.896	0.832	0.744	0.803	0.693
	Qwen-3-14B-it					
	First concept vector	0.980	0.936	0.910	0.900	0.920
	Mean of 5 concept vectors	0.969	0.920	0.891	0.858	0.926
	Linear probing					
	Llama-3.1-8B					
	Last token	**0.985**	0.9445 ± 0.0102	0.916	0.913	0.919
	Mean pooling	0.983	0.9435 ± 0.0088	0.914	0.911	0.918
	Max pooling	0.983	0.9470 ± 0.0099	0.920	0.911	0.930
	Qwen-3-8B-it					
	Last token	0.983	**0.9505** ± 0.0087	**0.925**	**0.920**	0.930
	Mean pooling	0.982	0.9420 ± 0.0134	0.912	0.914	0.909
	Max pooling	0.981	0.9500 ± 0.0083	**0.925**	0.914	**0.936**
	Qwen-3-14B-it					
	Last token	**0.985**	0.9430 ± 0.0097	0.914	0.907	0.921
	Mean pooling	0.981	0.9430 ± 0.0111	0.914	0.910	0.918
	Max pooling	0.979	0.9480 ± 0.0082	0.922	0.910	0.935
2*Original paper	EnvRoBERTa	-	**0.9565** ± 0.0098	**0.9319** ± 0.0140	**0.9330** ± 0.0399	0.9331 ± 0.0314
	EnvDistilRoBERTa	-	0.9502 ± 0.0113	0.9235 ± 0.0165	0.9064 ± 0.0330	**0.9423** ± 0.0176

Table 4: Classification Results on the Social Pillar of the ESG Dataset. Within each method, the best value per metric is shown in bold.

Category	Method	AUC	ACC	F1	Pr	Re
2*Baselines	Embedding method	0.9768 ± 0.0039	0.9245 ± 0.0185	0.9093 ± 0.0210	0.8818 ± 0.0315	0.9391 ± 0.0208
21*Experiment	*RFM*					
	Gemma-4-31b-it					
	First concept vector	**0.942**	**0.886**	**0.862**	0.820	**0.908**
	Mean of 5 concept vectors	0.923	0.838	0.792	0.798	0.786
	Llama-3.1-8B					
	First concept vector	0.921	0.860	0.821	**0.825**	0.816
	Mean of 5 concept vectors	0.889	0.818	0.774	0.754	0.796
	Qwen-3-8B-it					
	First concept vector	0.907	0.834	0.779	0.816	0.745
	Mean of 5 concept vectors	0.827	0.754	0.725	0.645	0.827
	Qwen-3-14B-it					
	First concept vector	0.935	0.858	0.824	0.802	0.847
	Mean of 5 concept vectors	0.916	0.842	0.790	0.823	0.760
	Linear probing					
	Llama-3.1-8B					
	Last token	0.964	0.9040 ± 0.0124	0.880	0.887	0.873
	Mean pooling	0.968	0.9120 ± 0.0139	0.890	0.891	0.891
	Max pooling	0.972	0.9210 ± 0.0121	0.902	0.897	0.908
	Qwen-3-8B-it					
	Last token	0.963	0.9030 ± 0.0120	0.879	0.880	0.878
	Mean pooling	0.970	0.9110 ± 0.0142	0.889	0.892	0.886
	Max pooling	0.976	0.9220 ± 0.0141	0.902	0.907	0.898
	Qwen-3-14B-it					
	Last token	0.966	0.9045 ± 0.0114	0.880	0.887	0.874
	Mean pooling	0.969	0.9170 ± 0.0165	0.896	0.899	0.894
	Max pooling	**0.977**	**0.9240** ± 0.0178	**0.910**	**0.907**	**0.950**
2*Original paper	SocRoBERTa	-	**0.9341** ± 0.0140	**0.9190** ± 0.0179	0.9035 ± 0.0345	**0.9366** ± 0.0292
	SocDistilRoBERTa	-	0.9319 ± 0.0146	0.9124 ± 0.0186	**0.9125** ± 0.0271	0.9129 ± 0.0254

Table 5: Classification Results on the Governance Pillar of the ESG Dataset. Within each method, the best value per metric is shown in bold.

Category	Method	AUC	ACC	F1	Pr	Re
2*Baselines	Embedding method	0.9184 ± 0.0116	0.8355 ± 0.0072	0.7313 ± 0.0109	0.6532 ± 0.0216	0.8327 ± 0.0399
21*Experiment	*RFM*					
	Gemma-4-31b-it					
	First concept vector	0.870	**0.866**	**0.717**	0.752	0.685
	Mean of 5 concept vectors	**0.885**	0.856	0.710	0.710	**0.710**
	Llama-3.1-8B					
	First concept vector	0.877	0.852	0.702	0.702	0.702
	Mean of 5 concept vectors	0.879	0.856	0.621	**0.894**	0.476
	Qwen-3-8B-it					
	First concept vector	0.863	0.842	0.599	0.808	0.476
	Mean of 5 concept vectors	0.846	0.822	0.590	0.688	0.516
	Qwen-3-14B-it					
	First concept vector	0.873	0.858	0.679	0.773	0.605
	Mean of 5 concept vectors	0.868	0.828	0.669	0.640	0.702
	Linear probing					
	Llama-3.1-8B					
	Last token	0.913	0.8595 ± 0.0089	0.718	0.781	0.665
	Mean pooling	0.918	0.8645 ± 0.0203	0.729	0.789	0.678
	Max pooling	0.908	0.8640 ± 0.0159	0.733	0.775	0.697
	Qwen-3-8B-it					
	Last token	0.907	0.8600 ± 0.0155	0.721	0.777	0.673
	Mean pooling	0.895	0.8460 ± 0.0136	0.685	0.763	0.624
	Max pooling	**0.923**	**0.8755** ± 0.0137	**0.757**	**0.795**	**0.723**
	Qwen-3-14B-it					
	Last token	0.907	0.8575 ± 0.0215	0.714	0.774	0.664
	Mean pooling	0.900	0.8590 ± 0.0155	0.713	0.786	0.654
	Max pooling	0.921	0.8720 ± 0.0251	0.749	0.791	0.712
2*Original paper	GovRoBERTa	-	0.8961 ± 0.0113	0.7848 ± 0.0262	**0.8562** ± 0.0184	0.7252 ± 0.0378
	GovDistilRoBERTa	-	**0.8966** ± 0.0111	**0.7886** ± 0.0159	0.8538 ± 0.0465	**0.7344** ± 0.0152

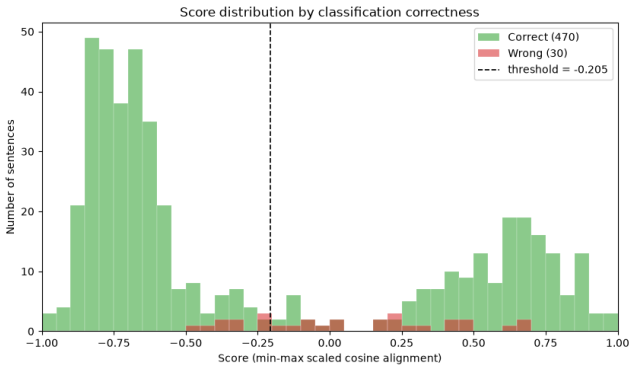
4.2 RFM continuous score

In Figure 1 we can see the distribution of the scaled continuous scores from the RFM mean of top 5 concept vectors method between -1 and 1 on all the ESG pillars. All methods had an optimised threshold to achieve the highest F1 score, everything left of the threshold is classified as not containing the concept and right of the threshold it is.

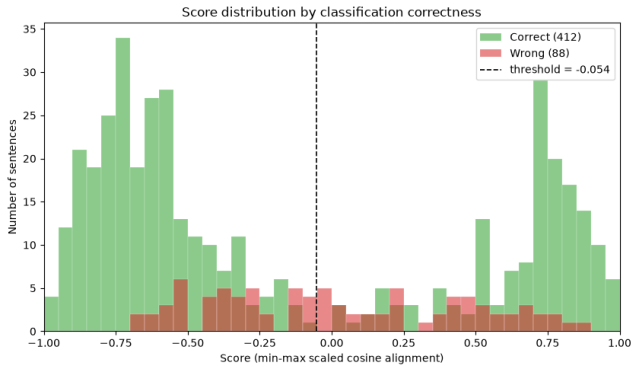
In Figure 1a we can see the distribution of Environmental scores, we can see that there is a separation on correctly labelled classes in green, with the label 0 scores clustering around -0.75 and the label 1 scores clustering around 0.65. Incorrectly labelled classes, coloured red, are scattered across the range with a small left skew.

Figure 1b has a clear separation in the correctly classified classes, shown in green, with label 0 scores clustering around -0.70 and label 1 scores clustering around 0.75. Incorrectly classified scores, shown in red, appear to have a small bell curve distribution around 0.

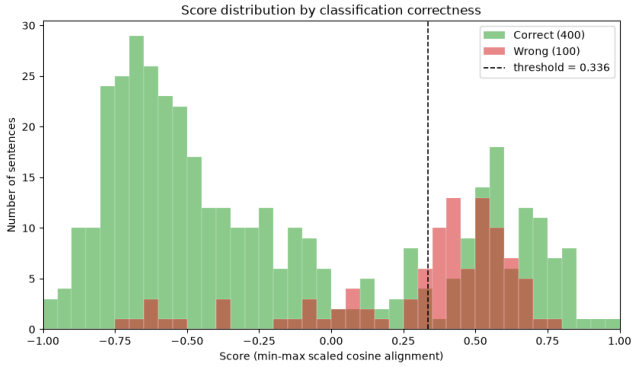
Then Figure 1c shows the distribution of Governance scores, here, there is no clear separation on the correctly classified data which is shown in green. The incorrectly classified data, shown in red, is predominantly clustered above the label 1 threshold.



(a) Environmental



(b) Social



(c) Governance

Figure 1: Distributions of the continuous RFM scores (mean of the top 5 concept vectors), scaled to [-1, 1] for each ESG pillar, grouped by classification correctness. Green denotes correctly classified sentences and red denotes misclassified ones; the dashed line marks the decision threshold.

5 Discussion

This thesis asked whether the gap between what an LLM knows internally and what it expresses in its output can be used to analyse financial text. The answer is yes: monitoring the internal activations of frozen, out-of-the-box models classifies ESG content close to fully fine-tuned models, without any fine-tuning. Our central expectation, however, was not met. The RFM concept-vector method did not outperform existing methods as it did in Beaglehole, Radhakrishnan, et al., 2026 findings, nor did it beat the fine-tuned models of Schimanski et al., 2024. In our domain, the simpler linear probe was consistently the stronger activation-based method.

Linear probing came closest to the fine-tuned models of Schimanski et al., 2024, but the best-performing pooling strategy was not consistent across pillars. On the Environmental task, last-token pooling performed best, while on the Social and Governance tasks max pooling was the strongest. One possible explanation is that Environmental is the easiest task, as all methods achieve high results on it, so the last token alone may already carry a rich enough summary of the sentence. The other two tasks appear more difficult, and max pooling may help by capturing the

few strong signals that are present more effectively than the other strategies.

The RFM method underperformed both linear probing and the fine-tuned models, and on the Social task it even fell below the embedding baseline. Within RFM, however, we found a clear and consistent pattern: larger models perform better than smaller ones, which is consistent with Beaglehole, Radhakrishnan, et al., 2026 findings, and the first concept vector matched or outperformed the mean of the top five concept vectors on accuracy and F1 across every pillar. This is contrary to what we expected, as the averaging of several directions should capture more of a concept that is spread across the activation space. A possible explanation is that the first concept vector, being the top eigenvector of the AGOP matrix, already captures the concept, while the additional vectors capture more peripheral concepts which adds noise.

Beyond classification, the RFM method provides a continuous score that indicates how strongly a concept is present in a new sentence. The distributions of these scores (Figure 1) separate the correctly classified sentences most clearly on the Environmental pillar and least clearly on Governance, which matches the accuracy achieved on each pillar. We also observe that the optimal threshold is never exactly 0. A threshold of 0 would indicate a perfect concept vector, since any genuine similarity would yield a positive score and any non-similarity a negative one. In our case the threshold was never 0, which suggests the concept vectors are not perfectly aligned with the concept; one possible interpretation is that they carry a general financial signal, since all sentences in the dataset are finance-related.

The misclassified scores behave differently across pillars. On Governance, almost all misclassified sentences are clustered between 0.25 and 0.50 and are mainly false positives, while on Social the misclassified scores form a small bell curve around 0, they sit between the correctly labelled sentences which suggests the similarity is uncertain on these sentences rather than confidently wrong. However what we found was that when the method does confidently misclassify a sentence, that sentence may be mislabelled by the original annotators. For example, using Gemma-4-31b-it on the Environmental task, the method confidently scored this sentence: "We are committed to creating a successful, sustainable business that is economically, socially and environmentally conscious." in the high range of other positive scores, even though the dataset labels it negative. We explore this further in Appendix B. We can however not confirm this for certain without re-annotation by experts, but it could be possible that some of the confident false- positive and negatives may be because the labels, rather than the method, is wrong.

Finally, one of the more surprising findings was the sensitivity across all methods to the wrapper used. Small changes, such as ending the wrapper with "Final answer:" versus "Answer:", already have a noticeable impact on the scores achieved by the methods (Appendix A). Optimising the wrapper therefore proved to be one of the more important steps in achieving good results, most likely because it guides the model towards the concept. The wrapper is also not model- or task-agnostic: the same wrapper applied to different models gives different results relative to each model's performance. One wrapper may perform better on Llama while another performs better on Qwen. A possible explanation is that this depends on what each model has already learned, some models may have seen more financial ESG data during training than others, so a wrapper that explains the concept in more detail may help a model that knows less about the topic, while confusing one that already knows it well.

5.1 Contributions

This thesis makes three main contributions. First, we show that activation-based methods provide a cheap, off-the-shelf measure of ESG content that comes close in performance to fully fine-tuned models. Both linear probing and RFM operate on frozen, out-of-the-box models and require no fine-tuning, yet linear probing comes within a few percentage points of the fine-tuned models of Schimanski et al., 2024.

Second, we present a continuous score, derived from the RFM concept vectors, that is intended to reflect how strongly a concept is present in a new sentence rather than only classifying it.

Third, we find that the choice of wrapper is one of the most important factors in obtaining a good concept vector. Small changes to the wrapper noticeably change the quality of the extracted concept vectors.

5.2 Practical implications

Because our activation-based methods come close to the fully fine-tuned models of Schimanski et al., 2024 while being based on frozen, out-of-the-box LLMs, they offer a valid option for anyone who does not have enough data or computing power to fine-tune a model themselves. The main saving will be in the gathering of data and training time, as with our method there is no domain-specific fine-tuning required. Although extracting the activations of an LLM still has a computing cost and there is a need for a small amount of data in order to extract the desired concept-vector.

From our activation-based methods, the simpler linear probe has consistently outperformed RFM, so the lightweight probe is worth considering before the slightly more complex concept vector method. That said, the continuous RFM score lets a text be ranked by how similar it is to the concept vector, and is intended to reflect how present a concept is rather than only whether it is present. An important note is that our results do show that choosing a wrapper has a significant impact on the outcome, since small changes to it noticeably affect the quality of the extracted concept vectors.

5.3 Limitations

Due to the methods being so sensitive to the wrapper, it could entirely be plausible that for concept X, a wrapper Y exists that would allow the RFM algorithm to capture the concept far more precisely than we managed, possibly achieving near-perfect separation on the data.

This further limits the RFM method itself, which depends entirely on the quality of the found concept vectors. If the direction is weak, so is the concept vector, and thus the monitoring. The linear probe does not suffer from this limitation of relying on a strong enough direction, it only has to find a boundary between the activations, which can be minimal.

This makes it difficult to separate how much of RFM's weaker performance is due to the method's dependence on a strong direction, and how much is due to us not finding the optimal wrapper.

In addition to that, ESG ratings are a niche subset within the financial sector. Our small models have likely seen a limited amount of ESG corpus during training, which may directly result in models having a less well-defined direction that captures each of the ESG pillars. We see support for this directly in the results: Environmental sentences consistently score higher across

all methods than Governance sentences. This is likely due to Environmental topics coming up more often in textual data than the specifics of governance structures. Using larger models, or a financially fine-tuned model like the base models of Schimanski et al., 2024, would allow stronger directions of the ESG pillars to be present.

We also had issues with using the linear probe on Gemma-4-31b-it, despite our efforts the kernel kept running out of memory before we could extract any result, thus it being absent from our tables.

Lastly, fully validating the continuous RFM score as a measure of how much a concept is present, rather than only whether it is, is not yet possible in our domain. Current ESG datasets, including the one we use (Schimanski et al., 2024), have only binary labels, so we leave a direct validation of the score's magnitude to future work.

6 Conclusion

This study asked whether the gap between what an LLM knows internally and what it expresses in its output can be used to better analyse financial text. We found the answer to be yes, with the use of frozen, out-of-the-box LLMs we successfully classified ESG content outperforming our baseline in most cases and coming close to fully domain-specific fine-tuned models, without us fine-tuning the LLMs at all. The RFM method, however, did not outperform existing methods as the simpler linear probe consistently outperformed it. We also found that the choice of wrapper has a large impact on the quality of the extracted concept vectors. The RFM method does provide a continuous score that is intended to reflect how strongly a concept is present in a text, rather than only whether it is present. Future research could, first, develop a more reliable way to design the wrapper, so that it best guides the model towards the desired concept vector. And second, validate whether the continuous RFM score truly reflects how strongly a concept is present in a new sentence.

References

Alain, G., & Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. https://openreview.net/forum?id=HJ4-rAVtl

Azaria, A., & Mitchell, T. (2023, December). The internal state of an LLM knows when it's lying. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 967–976). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.68

Beaglehole, D., Holzmüller, D., Radhakrishnan, A., & Belkin, M. (2026). XRFM: Accurate, scalable, and interpretable feature learning models for tabular data. *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=wHuVdpnUFp

Beaglehole, D., Radhakrishnan, A., Boix-Adserà, E., & Belkin, M. (2026). Toward universal steering and monitoring of ai models. *Science*, *391*(6787), 787–792. https://doi.org/10.1126/science.aea6792

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, *48*(1), 207–219. https://doi.org/10.1162/coli_a_00422

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2023). Discovering latent knowledge in language models without supervision. *The Eleventh International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=ETKGuby0hcs

Du, K., Zhao, Y., Mao, R., Xing, F., & Cambria, E. (2025). Natural language processing in finance: A survey. *Information Fusion*, *115*, 102755. https://doi.org/https://doi.org/10.1016/j.inffus.2024.102755

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Hewitt, J., & Liang, P. (2019, November). Designing and interpreting probes with control tasks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 2733–2743). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1275

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Marks, S., & Tegmark, M. (2024). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Conference on Language Modeling (COLM)*. https://openreview.net/forum?id=aajyHYjjsk

Park, K., Choe, Y. J., & Veitch, V. (2024, July). The linear representation hypothesis and the geometry of large language models [Conference dates: 21–27 Jul]. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st international conference on machine learning* (pp. 39643–39666, Vol. 235). PMLR. https://proceedings.mlr.press/v235/park24c.html

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. *Finance Research Letters*, *61*, 104979.

Tenney, I., Das, D., & Pavlick, E. (2019, July). BERT rediscovers the classical NLP pipeline. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4593–4601). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1452

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter brian, b., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 24824–24837, Vol. 35). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., ... Hendrycks, D. (2023). Representation engineering: A top-down approach to ai transparency.

Appendix

A Wrappers

The quality of the activations is highly dependent on the wrapper that is used on the data. For example there is a noticeable difference in the results of a wrapper ending with "answer: " instead of "final answer: ". There also is a balance between how much context you should give the model, over explaining a concept can result into achieving a worse result. Table 6 highlights the differences between the results based on what wrapper is used. All other variable remain the same, we use the RFM algorithm using Gemma-4-31b, on the Governance dataset. {statement} is used as placeholder for the actual sentence used.

Table 6: Wrapper sensitivity on the Governance dataset (RFM, Gemma), using the averaged score across all five concept vectors, evaluated on the same held-out test split. Best value per metric in bold.

Wrapper	Threshold	AUC	ACC	F1
{statement}	+0.240	0.768	0.784	0.530
Does the following sentence **discuss** corporate governance? Sentence: {statement} Final answer:	+0.506	0.735	0.766	0.354
Text: {statement} Does this text **discuss** Governance as accountability or oversight — such as board structures, ethics codes, anti-corruption, transparency in decision-making, or regulatory compliance? **Let's think step by step.** Final answer:	−0.087	0.842	0.802	0.655
Text: {statement} Does this text **discuss** Governance as accountability or oversight — such as board structures, ethics codes, anti-corruption, transparency in decision-making, or regulatory compliance? **Reason step-by-step. Answer:**	+0.350	**0.911**	**0.860**	**0.733**
Text: {statement} Does this text **address** Governance as accountability or oversight — such as board structures, ethics codes, anti-corruption, transparency in decision-making, or regulatory compliance? **Reason step-by-step.** Final answer:	+0.251	0.870	0.802	0.650
Text: {statement} Does this text **discuss** Governance as accountability or oversight — such as board structures, ethics codes, anti-corruption, transparency in decision-making, or regulatory compliance? **Reason step-by-step.** Final answer:	+0.169	0.881	0.834	0.684

B Confidently misclassified RFM continuous scores

The RFM method provides us with a score of how similar new activations are compared to the concept vector. We scale the results between -1 and 1 in order to have a clear continuous score, this score can be interpreted as to how much a concept is active in a new sentence and how confident the model is. Analysing these results show that the model sometimes very confidently misclassified sentences from the ESG dataset created with three human annotators by Schimanski

et al., 2024. However the inter-annotator agreement is never 100% on the datasets, and the lowest agreement is on the dataset that has proven to be the most difficult to score accurately by all models, the Governance dataset. Yet when we look at the continuous scores our RFM method provides, the confident errors lie on genuinely boundary-line cases. If we look at Figure 1, we can see the distribution of the scores. Now when we look at the top 5 most confidently misclassified scores, shown in Table 7 we can see that some of the false positive sentences do mention governance topics, as well as these topics being absent in some false negative cases. However a new team of domain experts should re-annotate these sentences in order to make real claims.

Table 7: Top 5 confidently misclassified Governance sentences (RFM, Gemma).

Score	Sentence
False positives (Pred = 1, Label = 0)	
+0.713	This includes the submission of an annual report on our policies and procedures and the development of an online training program for all employees.
+0.693	Tenure, election, reappointment and removal of Directors Directors are typically appointed by the Board and then put forward for election by shareholders at the subsequent AGM.
+0.692	The firm's mandatory all-employee training, which was introduced for Board members in Q4 2021, included sessions on Customer Vulnerability.
+0.622	The Board of Directors of the Company consists of eight directors.
+0.621	In addition, the Board has introduced a policy to ensure that the financial and social performance of each of the Group's service lines and regions is reported in the report.
False negatives (Pred = 0, Label = 1)	
−0.479	These are available to our analysts and investment teams to help them identify, understand and assess climate risks for different types of assets, drawing on a database of ESG information refreshed daily by our data vendors.
−0.431	transition risks, emerging from national and global policy responses to the climate crisis (such as regulatory changes, taxes and levies, etc) and the transition to a sustainable climate model (such as changes in energy availability and mix, disruptions to the company's own or partner business models, and reduced availability of unsustainable components or materials).
−0.297	We aim to get the balance right between short-term delivery and long-term sustainability, between top-line growth and overall stakeholder value creation.
−0.244	Accounting for sales, grant income and deferred income relating to Vaxzevria (Group) Refer to Audit Committee Report, Group Accounting Policies and Notes 20 in the Group Financial Statements In 2020, the Group entered into an arrangement with the University of Oxford for the global development, production and supply of the COVID-19 vaccine, Vaxzevria.
−0.238	We have been a member of the FTSE4Good index since 2001 and the FTSE4Good Environmental Leaders Europe index since 2001.