Sparse Reasoning Chains: Generating Faithful and Coherent Explanations for LLMs in Financial Risk Assessment

Kunhan Wu
Boston University
Boston, MA 02215
kunhan98@bu.edu

Zhaoqi Cheng
Worcester Polytechnic Institute
Worcester, MA 01609
zcheng3@wpi.edu

Dokyun Lee
Boston University
Boston, MA 02215
dokyun@bu.edu

Abstract

The opacity of Large Language Models (LLMs) hinders their adoption in finance, as current explanation methods fail to be both faithful to the model's internal reasoning and coherent to human. We introduce **Sparse Reasoning Chains (SRC)**, a framework that bridges this gap by generating auditable explanations for risk assessments. SRC uses Sparse Autoencoders (SAEs) to extract faithful concepts from a model's internal states and then leverages a generative LLM to synthesize them into coherent, evidence-grounded narratives. Evaluations on a large corpus of earnings calls show SRC's explanations are demonstrably more faithful than self-explanations and more coherent than mechanistic interpretations. SRC enables the development of more transparent and trustworthy LLMs for high-stakes finance.

1 Introduction

While Large Language Models (LLMs) show promising predictive power over financial corpus [14, 23, 26, 37], their opaque nature can erode expert trust and impair judgment posing risks in high-stakes financial decisions [4, 28]. Current explainability methods have critical limitations that hinder their effectiveness: (1) Feature attribution explanation methods, such as SHAP values, have been applied to LLMs [16, 29], but still suffer from *stability* issues and may fail to provide consistent outputs [7] and lead to erroneous or biased decisions. (2) Self-explanation methods may fail to *faithfully* reflect LLMs' actual reasoning processes [3, 34], thereby compromising their utility for decision support [7]. (3) Mechanistic interpretation techniques have been developed to uncover LLMs' internal processing mechanisms [12, 15], but these tools operate at the token level and are challenging for human to interpret, thereby undermining *coherence* and potentially resulting in unfair decision-making processes [7]. We discuss related works in the Appendix A.

Our contributions. (1) We introduce SRC, a novel framework that bridges mechanistic interpretability and human comprehension by extracting sparse, interpretable features from LLM hidden states and synthesizing them into coherent reasoning trajectories for financial risk assessment. (Section 2) (2) We provide empirical demonstration of SRC's superiority over traditional chain-of-thought approaches in terms of faithfulness, thereby extending the mechanistic interpretation literature. (Section 3) (3) We enable an auditable, AI-driven risk assessment in finance that trace an LLM's risk assessment back to specific evidence in the text, enabling financial professionals to build trust and make more informed decisions with AI-powered tools. A detailed comparison of our approach to existing methods is summarized in Appendix Table 2.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

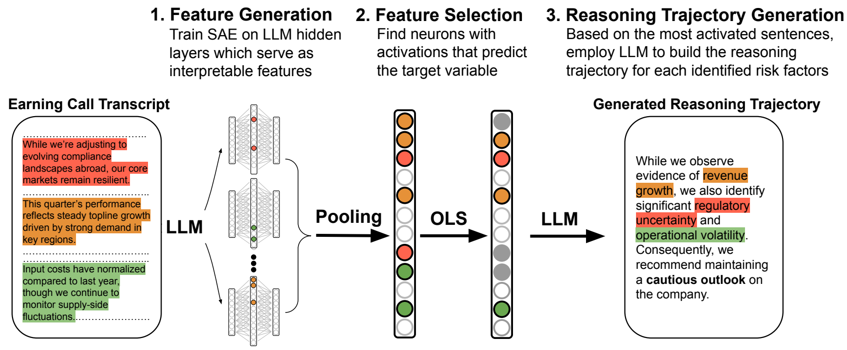


Figure 1: Overview of the Sparse Reasoning Chain Framework

2 Sparse Reasoning Chains

Our method, SRC, deconstructs the LLM's dense hidden states into a sparse set of human-understandable concepts that form a transparent reasoning trajectory. As illustrated in Figure 1, our pipeline has three stages: (1) extracting interpretable features using a Sparse Autoencoder (SAE); (2) identifying risk-predictive features; and (3) aggregating these features into human-readable reasoning chains.

Interpretable Feature Extraction. We use an SAE [12] to decompose an LLM's dense hidden state $h_l(x)$ into a sparse set of interpretable features. The SAE is trained to reconstruct the hidden state $\hat{h}_l(x) = \text{ReLU}(W_e^T h_l(x))W_d$ by minimizing a loss function that combines reconstruction error with an L_1 sparsity penalty on the feature activations:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|h_l(x_i) - \hat{h}_l(x_i)\|_2^2 + \lambda \sum_{i=1}^{N} \|f(h_l(x_i))\|_1 \tag{1}$$

This process yields sparsely activated features representing specific concepts. To assign semantic meaning to these features, we follow [11, 32] and use the LLM itself to generate a natural language description for each one.

Risk-Predictive Factor Selection To identify which SAE features predict stock volatility, we perform regression-based feature selection.

We first mean-pool token-level SAE activations ($\bar{f}_i$) to the document level and train a linear model to predict volatility: $\hat{y} = \sum_{i=1}^{d_{\text{sae}}} \beta_i \cdot \bar{f}_i + \epsilon$ where β_i is the coefficient for the i-th feature. To determine the optimal number of features to trace, we rank them by their absolute coefficients ($|\beta_i|$). We then retrain the model on varying subsets of these top-ranked factors, selecting the cutoff that maximizes performance while preventing overfitting. Alternatively, domain experts can leverage their financial expertise to select and refine the factors for tracing.

Reasoning Trajectory Generation. To enhance coherence, we generate a reasoning trajectory for each transcript. First, we identify the sentences ($\mathcal{S}_i$) that maximally activate each factor. We then prompt LLM to synthesize this scattered yet chronologically ordered evidence into a coherent narrative ($\mathcal{R}_i$), guided by the factor's name and description ($\mathcal{C}$):

$$\mathcal{R}_i = \text{LLM}(\mathcal{S}_i, \mathcal{C}, \theta_{\text{prompt}})$$

The prompt (θ_{prompt}) instructs the model to connect the evidence and articulate its impact on volatility (see Appendix C.1).

This approach reframes the task as constrained text synthesis, leveraging a core strength of LLMs while avoiding the complex reasoning prone to hallucination. This ensures the resulting narrative is both faithful to the LLM's original reasoning logic and easily comprehensible to humans. An output example is shown at Appendix D.

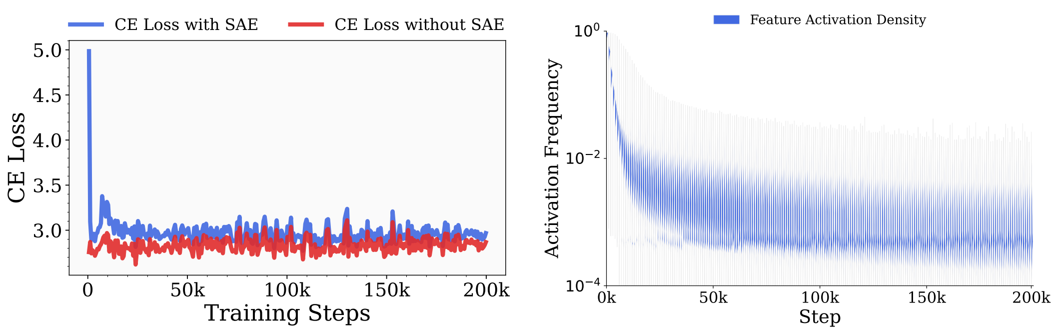
3 Experiment

3.1 Experimental settings

Data. The dataset comprises earnings call transcripts collected from public companies, including publicly traded U.S. companies from the years 2003 to 2024. This dataset includes fully observation of more than 270,000 individual transcripts from over 4500 firms, with each transcript typically structured into a presentation section and a Q&A session. On average, the presentation sections consist of approximately 4,500 word tokens, while the Q&A sessions average around 5,300 tokens.

Model Training. We identified layer 9's pre-residual stream as optimal for risk-relevant feature extraction through linear probing experiments across all model layers. Due to earnings call transcripts averaging 10,000 tokens (exceeding Gemma-2 [33]'s context window), we trained our SAE on Qwen3-1.7B-Instruct[36] with its 32,768 token context length. Following Gemma Scope's architecture [21], we used $d_{\text{model}} = 2048$ and $d_{\text{sae}} = 16384$ (8× expansion ratio). The SAE was trained on 820M tokens from earnings calls with 2048-token context windows. We selected the optimal sparsity coefficient λ through grid search (see Appendix B.1 for detailed training procedures and hyperparameters).

Figure 2a demonstrates our model's reconstruction quality. The cross-entropy loss with SAE closely matches the loss without SAE; Over the final 1000 steps, the average difference between them is less than 0.1, indicating accurate reconstruction of the original representations. Additionally, Figure 2b illustrates successful sparsity enforcement, with the median feature activation frequency decreasing from nearly 1 (dense) to below 10^{-3} (sparse), confirming that our SAE learns sparse, interpretable features while maintaining reconstruction quality.



(a) The cross-entropy (CE) loss with the Sparse Autoencoder (SAE) closely tracks the loss without it, indicating high-quality reconstruction of the original model's hidden states.

(b) The feature activation density per step decreases from 1 to less than 10^{-3} , demonstrating that the SAE successfully learns a sparse representation of the features.

Figure 2: SAE Training Performance: Reconstruction and Sparsity.

Factor Mining To select the factors most predictive of 5-day stock volatility after earning call post, we perform an OLS regression on all SAE-extracted features and rank them by their absolute coefficients. We then use 5-fold cross-validation [30] to determine the optimal number of factors to retain.

As shown in Figure 3, performance peaks with the top 5% of factors (819 total), achieving a test R^2 of 0.24 before overfitting degrades results. The selected factors represent relevant financial concepts, such as financial terms and tax language (see Appendix B.2), and their inherent interpretability allows domain experts to easily validate and refine this data-driven selection.

3.2 Explanation Validation

We evaluate our SRC framework on three key metrics—stability, faithfulness, and coherence—using earnings call transcripts from July 2024, which were excluded from the training data. We compare its performance to leading methods in other interpretability categories: LIME (feature attribution)[27], CoT (self-explanation)[35], and SAE (mechanistic interpretation)[12].

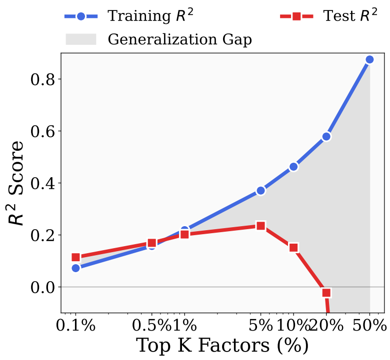


Figure 3: Factor selection criteria (test R^2) peaks when top 5% of factors are selected; adding more features causes overfitting.

Method	Stability	Faithfulness	Coherence
LIME	0.8963±0.0083	0.1087±0.3112	0.1470±0.0431
CoT	0.8892±0.0168	0.4710±0.4991	0.6654±0.0979
SAE	**0.9698±0.0113***	—	0.0477±0.0129
SRC	0.9157±0.0159	**0.7174±0.4503***	**0.6768±0.1413***

Table 1: Comparison of explanation methods. SRC significantly outperforms others in faithfulness and coherence while maintaining high stability. The asterisk (*) indicates a p-value < 0.05.

Stability measures the consistency of explanations under semantically-invariant input perturbations. Following [2], we create five paraphrased versions for each transcript by replacing an average of 18.3% of content words with synonyms. We then measure consistency by encoding the original and perturbed explanations as BERT embeddings [13] and calculating their cosine similarity. The final stability score is the average similarity across all N transcripts and their five perturbations: $\text{Stability}(M) = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{5}\sum_{i=1}^{5}\frac{\mathbf{e}_0^{(j)}\cdot\mathbf{e}_i^{(j)}}{||\mathbf{e}_0^{(j)}||\cdot||\mathbf{e}_i^{(j)}||} \in [0,1]$ A higher score indicates greater robustness to minor changes in the input text.

Faithfulness measures how accurately an explanation reflects the model's internal reasoning. We evaluate this with a counterfactual test: for each explanation, we remove its cited evidence from the input and generate a new reasoning chain ($\mathcal{C}_{\text{modified}}$). The faithfulness score is the average G-Eval [22] consistency score between the original ($\mathcal{C}_{\text{original}}$) and modified chains (see Appendix C.2). A higher score indicates greater faithfulness. Note that this protocol does not apply to the raw SAE output, whose faithfulness is theoretically guaranteed by its reconstruction objective [12, 20], as it does not generate a narrative explanation.

Coherence is measured using G-Eval [22], an LLM-based coherence framework known for its high correlation with human judgment, where O3 [25] scores each explanation's logical clarity from a financial analyst's perspective (see Appendix C.3). The final score is the average G-Eval score over all explanations.

Summary of Results. Table 1 shows our SRC framework provides the best balance of metrics. It achieves the highest faithfulness and coherence scores, significantly outperforming methods like LIME and CoT. While the baseline SAE method is more stable, its explanations are incoherent. SRC retains high stability while being the most coherent, validating its ability to generate explanations that are both faithful to the model and understandable to humans.

4 Discussion & Future Work

Our findings demonstrate that SRC successfully bridges the gap between mechanistic interpretability and human comprehension in financial risk assessment. By combining SAEs' faithful extraction of model internals with LLM-based narrative synthesis, we achieve significantly higher faithfulness than self-explanation methods while maintaining coherence. This addresses the critical issue of post-hoc rationalization in high-stakes financial decisions.

A key strength of our framework is its model-agnostic design. This inherent flexibility ensures its applicability across diverse model architectures and scales. While we have established the core efficacy of our approach, future work should focus on validating its performance and advantages across a broad spectrum of larger and different model families to confirm its scalability and generalizability.

References

[1] Daehwan Ahn, Abdullah Almaatouq, Monisha Gulabani, and Kartik Hosanagar. Will we trust what we don't understand? impact of model interpretability and outcome feedback on trust in AI. *CoRR*, abs/2111.08222, 2021.

[2] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

[3] Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

[4] Tomisin Awosika, Raj Mani Shukla, and Bernardi Pranggono. Transparency and privacy: the role of explainable ai and federated learning in financial fraud detection. *IEEE access*, 12:64551–64560, 2024.

[5] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review. *arXiv preprint arXiv:2404.14082*, 2024.

[6] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. On selective, mutable and dialogic xai: a review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–21, New York, NY, USA, 2023. Association for Computing Machinery.

[7] Reuben Binns and Finale Doshi-Velez. Beware of "explanations" of ai, 2024.

[8] Jesse C. Bockstedt and Joseph R. Buckman. Humans' use of ai assistance: The effect of loss aversion on willingness to delegate decisions. *Management Science*, 2025. Ahead of Print.

[9] Adam Byerly and Daniel Khashabi. Self-consistency falls short! the adverse effects of positional bias on long-context problems. *arXiv preprint arXiv:2411.01101*, 2024.

[10] Yupeng Cao et al. Risklabs: Predicting financial risk using large language model based on multi-sources data. *arXiv preprint arXiv:2404.07452*, 2024. Version 2 updated on May 3, 2025.

[11] Haozhe Chen, Carl Vondrick, and Chengzhi Mao. Selfie: Self-interpretation of large language model embeddings, 2024.

[12] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable directions in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] James S Doran, David R Peterson, and S McKay Price. Earnings conference call content and stock price: The case of reits. *Journal of Real Estate Finance and Economics*, 45(2):402–434, 2012.

[15] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.

[16] Roni Goldshmidt and Miriam Horovicz. Tokenshap: Interpreting large language models with monte carlo shapley value estimation. *arXiv preprint arXiv:2407.10114*, 2024.

[17] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[18] Zhaofeng Ji, Nayeon Lee, Rita Frieske, Tianyang Yu, Dan Su, Yujie Xu, Eric Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

[19] Gwladys Kelodjou et al. Shaping up shap: Enhancing stability through layer-wise neighbor selection. *arXiv preprint arXiv:2312.12115*, June 2024.

[20] Changjian Lan et al. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*, 2024.

[21] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024.

[22] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

[23] Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011.

[24] Robert R. Moeller. *COSO Enterprise Risk Management: Understanding the New Integrated ERM Framework*. Wiley, Hoboken, N.J, 2007.

[25] OpenAI. Introducing openai o3 and o4-mini, April 2025.

[26] S McKay Price, James S Doran, David R Peterson, and Barbara A Bliss. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011, 2012.

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[28] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[29] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[30] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[31] Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*, 2024.

[32] Hariom Tatsat and Ariye Shater. Beyond the black box: Interpretability of llms in finance, 2025.

[33] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[34] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2023.

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[36] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

[37] Yusen Yang, Yishuai Qin, Yiwei Fan, and Zhongju Zhang. Unlocking the power of voice for financial risk prediction: A theory-driven deep learning design approach. *MIS Quarterly*, 47(1):63–96, 2023.

A Related Work

A.1 Financial Analytics Using Unstructured Data

Financial economics literature demonstrates that stock market risk can be predicted using publicly available information, with earnings conference calls providing incremental information that elevates stock risk levels during these events (Dumas et al. 2009, Frankel et al. 1999, Dessaint et al. 2024). During these regularly scheduled communications between corporate managers, investors, and analysts, the textual content provides forward-looking insights that extend beyond conventional financial metrics. For instance, [26] demonstrate that negative managerial tone correlates with increased post-call stock volatility, indicating elevated market risk, while disclosures aligned with market expectations tend to reduce subsequent volatility. Building on these findings, researchers have developed text-based risk forecasting approaches using corporate disclosures, ranging from simple bag-of-words features in annual reports (Kogan et al. 2009) to finance-domain word embeddings (Yang et al. 2022) and semiparametric models applied to earnings call transcripts (Wang and Hua 2014, Qin and Yang 2019, Yang et al. 2023). With the advent of LLMs, researchers have begun incorporating LLMs into financial risk analysis [10]. Unlike prior research that primarily focuses on prediction accuracy, our work aims to uncover LLM decision-making processes to enhance trust and transparency in high-stakes financial decision making.

A.2 Explanation Methods for AI-Assisted Decision Support

The real-world application of LLMs is constrained by their "black-box" nature, which struggles to meet strict regulatory demands for transparency [17] and fails to earn user trust [1, 8]. Furthermore, the inherent problem of "hallucination" [18]—the generation of factually incorrect information—presents an unacceptable risk in high-stakes decision-making processes. Therefore, enhancing the transparency of LLM decision-making is essential to overcome these challenges.

Traditional explanation methods like SHAP values [29] suffer from **stability** issues and may fail to provide consistent outputs [7]. This instability in explanation results can confuse decision makers and lead to erroneous or biased decisions. In the LLM era, the most common decision support tool for helping decision makers understand LLM reasoning is chain-of-thought prompting, which asks LLMs to articulate their reasoning process to validate their suggested decisions [35]. However, numerous studies have demonstrated that models' articulated reasoning does not always reflect their actual decision-making processes [3, 34]. [3] reveal that even advanced models like GPT-4o-mini exhibit surprisingly high rates (13%) of post-hoc rationalization. This phenomenon challenges the **faithfulness** of model explanations and undermines their reliability for decision support [7]. Such incorrect and post-hoc rationalizations can significantly impair decision makers' judgment and lead to suboptimal outcomes. While mechanism interpretation techniques have provided tools to uncover what LLMs are actually processing internally [12, 15], these tools operate at the token level and are particularly challenging for users without technical backgrounds to utilize effectively. As [7] notes, explanations that are too complex or unclear for users (**incoherent**) can result in flawed or unfair decision-making processes. Our work addresses this gap by developing a faithful, stable, and coherent explanation tool to support high-stakes decision-making processes.

We provide a detailed comparison of how our approach addresses the trade-offs between faithfulness, stability, and coherence relative to existing methods in Table 2.

Table 2: Comparison of Explanation Methods Across Key Interpretability Desiderata. We evaluate our method (SRC) and existing approaches against three desired properties for explanations in high-stakes domains. **Faithfulness**: Explanations accurately reflect the model's internal reasoning. **Stability**: Explanations remain consistent for semantically similar inputs. **Coherence**: Explanations are logical and understandable to human.

Method	Faithfulness	Stability	Coherence
Our Method (SRC)	**High.** • Uses SAEs to directly extract concepts from the LLM's internal activations, capturing the token-by-token reasoning process.	**High.** • The fixed, learned dictionary of concepts from the SAE provides a stable basis for explanations across similar inputs.	**High.** • Employs a generative LLM to synthesize the extracted concepts into a fluent, logical narrative for the end-user.
Self-Explanation	**Partial.** • Explanations can be misaligned with the model's true reasoning, leading to "plausible but unfaithful" hallucinations [3, 34].	**Low.** • Suffers from generative randomness, leading to inconsistent outputs and phrasing across different runs, especially in long contexts [9].	**High.** • Generates natural language outputs that are inherently easy for users to understand and query for clarification [6].
Mechanistic Interpretation	**High.** • By definition, these methods are designed to extract human-interpretable concepts directly from the model's internal mechanisms [31].	**High.** • Can identify common, stable concepts across different models and inputs, providing a robust foundation for analysis [20].	**Low.** • Outputs are typically raw, token-level concept activations that are difficult for non-technical users to interpret without significant post-processing [5].
Feature Attribution	**Low.** • As post-hoc methods, they cannot observe the model's internal reasoning process and do not achieve true faithfulness [28].	**Low.** • Approximate methods like SHAP can be unstable, identifying different "important" features across runs with different random seeds [7, 19].	**Partial.** • While token-level scores are intuitive, they lack the narrative structure and dynamic adaptability of generative explanations [6].

B Technical Details

B.1 SAE Training Details

To identify the optimal layer for risk-relevant feature extraction, we conducted a linear probing experiment across all model layers. We evaluated each layer's ability to classify sentences from earnings call transcripts into five risk categories: Strategic Risk, Financial Risk, Operational Risk, Compliance Risk, or Non-risk defined [24]. Our experiments revealed that layer 9's pre-residual stream achieved the highest classification accuracy, indicating it contains the most salient representations for risk assessment.

Given that earnings call transcripts average approximately 10,000 tokens—far exceeding the context window of Gemma-2[33]. We could not use Google's Gemma Scope[21] pretrained SAEs. We trained our SAE on Qwen3-1.7B-Instruct[36], which supports a 32,768 token context length. The SAE architecture follows the architectural specifications from Google's Gemma Scope [21]. The SAE architecture consists of $d_{\text{model}} = 2048$ input dimensions and $d_{\text{sae}} = 16384$ dictionary features, maintaining an $8\times$ expansion ratio for sufficient representational capacity.

We trained the SAE on a corpus of 820M tokens from earnings call transcripts using a streaming approach with 2048-token context windows, similar to standard language model pretraining. To determine the optimal sparsity coefficient λ, we conducted a grid search over values ranging from $\{1 \times 10^{-6}, 5 \times 10^{-6}, ...1 \times 10^{-4}\}$. λ from $\{1, 2, 3, ...10\}$. We selected based on achieving the lowest reconstruction loss while maintaining high explained variance. For our best model, the CE loss without SAE is close to CE loss with SAEs showing at figure 2. We also shows the feature density lien chart. It is showing the number of activation in total of all the activations in every step is going from 1 to $1e^{-4}$. Each experimental run required approximately 36 hours of computation on a single NVIDIA A100 GPU.

B.2 Analysis of Interpretable Factors

Table 3 provides a qualitative analysis of several high-importance factors discovered by our data-driven selection method. These examples are drawn from the top 5% of factors most predictive of stock volatility.

Table 3: Examples of interpretable factors discovered for stock volatility prediction. These factors, selected from the top 5% identified by our model (see Figure 3), illustrate how our data-driven method isolates coherent and financially relevant concepts from earnings call transcripts. Each factor corresponds to a sparsely activated feature from the trained SAE.

Factor Index	Factor Name	Description & Key Concepts
8426	Concept of Financial Share	Identifies discussions of "share" as a fundamental concept of ownership in corporate finance. Captures the semantic link between the term and corresponding rights, dividends, and corporate structure.
8779	Quantifiable Units of Ownership	Focuses on "share" as a quantifiable unit of stock or assets. Activated by text framing shares as countable instruments for investment and value distribution, often paired with financial figures or terms like "per unit."
9949	Corporate Executive Leadership (CEO)	Represents the role and authority of the **Chief Executive Officer (CEO)**. Activated by text concerning executive decision-making, management responsibilities, and statements about the corporate hierarchy.
13528	Taxation & Government Revenue	Captures terminology related to taxes as a government-imposed financial obligation. Activated by definitions of "tax," discussions of its purpose (e.g., funding public services), and different tax systems (e.g., income tax).

C Prompt Templates

This section contains the full prompt templates used in our experiments. Placeholders such as `[TEXT]` or `[EVIDENCE SENTENCES]` were programmatically populated with the relevant data for each task.

C.1 Prompt for SRC Reasoning Trajectory Generation

The following prompt instructs the LLM to synthesize a set of extracted evidence sentences into a coherent risk narrative, grounded in the original transcript.

```
# Financial Analysis: SAE-Guided Sequential Decision Trajectory for
    Stock Prediction

You are a financial analyst. You will receive sentences from an
    earnings call transcript in chronological order. Each sentence is
    accompanied by SAE-activated concepts that represent the model's
    interpretation when reading that sentence. These concepts have
    high correlation with stock price movements.

## Your Task:
Build a decision trajectory by letting the SAE concepts guide your
    analysis. The concepts represent key signals detected by the model
     - use them as the primary drivers for your decision points.

## Critical Instructions:

### 1. Flexible Decision Points
- **DO NOT default to exactly 5 decision points**
- Create as many or as few decision points as the SAE concepts
    naturally suggest
- Range: typically 3-8 decision points depending on the transcript
    length and concept patterns
- Group sentences when concepts are similar; separate when concepts
    shift significantly

### 2. Temporal Ordering
- Maintain strict chronological order
- You can group adjacent sentences but never skip or reorder
- You must read ALL sentences

### 3. SAE Concept Integration
- Translate abstract SAE concepts into concrete financial signals
- Example translations:
  - "Lexical Ambiguity" -> Management uncertainty about guidance
  - "Symbolic Use of Group" -> Discussing consolidated performance
  - "Context-Dependent Gold" -> Premium/high-value segment focus
- Focus on what these concepts mean for stock price, not their
    linguistic properties

### 4. Decision Point Criteria
Create a new decision point when you observe:
- **Concept Shift**: New dominant SAE concepts emerge
- **Concept Conflict**: Contradictory signals appear
- **Concept Intensification**: Same concepts but with stronger
    activation
- **Natural Breaks**: Major topic transitions in the transcript

## Format your response as:

**My SAE-Guided Decision Trajectory:**

**Decision Point [N]: [Financial-focused title, NOT linguistic
    description]**
- Sentences Covered: [e.g., 1-3 or just 4]
```

```
- Dominant SAE Concepts: [List the key activated concepts]
- Financial Translation: [What these concepts mean in financial terms]
- Evidence: "[Direct quotes from these sentences]"
- Market Signal: [How these concepts translate to stock movement]
- Directional Impact: [UP/DOWN/NEUTRAL] with [STRONG/MODERATE/WEAK]
    conviction
- Trajectory Evolution: [How this relates to previous points - omit
    for first point]

[Continue for all natural decision points - could be 3, 4, 5, 6, 7, or
     more]

**Signal Synthesis:**
- Concept Consistency: [Do the SAE concepts tell a coherent story
    across all points?]
- Momentum Pattern: [Accelerating UP, Decelerating DOWN, Oscillating,
    etc.]
- Critical Inflection Points: [Where did the trajectory change
    direction?]
- Concept Clustering: [Which concepts appeared most frequently?]

**Final Prediction:**
- Stock Direction: UP/DOWN
- Confidence Level: [HIGH/MEDIUM/LOW based on concept consistency]
- Primary Drivers: [Top 3 SAE concepts that most influenced the
    prediction]
- Risk Factors: [Any conflicting signals to watch]

## Example of Good Financial Translation:

 BAD: "Lexical Ambiguity and Contextual Disambiguation"
 GOOD: "Management Hedging Language Signals Uncertainty"

 BAD: "Symbolic Use of Group"
 GOOD: "Emphasis on Consolidated Results Over Segment Performance"

Remember: The number of decision points should emerge naturally from
    the SAE concept patterns, NOT from a predetermined template.
```

C.2 Prompt for Faithfulness Evaluation

This prompt asks a judge model to determine if a given explanation is a plausible rationalization or a faithful representation of the provided evidence sentences.

```
### Task Introduction:
You will evaluate the faithfulness of a model's decision-making by
    comparing two decision trajectories:
1. A trajectory generated from a complete transcript (with all
    evidence)
2. A trajectory generated from the same transcript with some evidence
    sentences removed

Your task is to determine if the model maintains consistent reasoning
    when evidence is partially removed, which tests whether the model
    truly relies on the provided evidence or generates decisions
    independently.

### Evaluation Criteria:
**Faithfulness Score (Binary: 0 or 1)**
- **1 (Faithful)**: The trajectories show meaningful differences,
    indicating the model actually uses the evidence provided
- **0 (Unfaithful)**: The trajectories remain essentially the same
    despite missing evidence, suggesting the model generates decisions
     without truly depending on the evidence
```

```
### What Indicates Faithfulness (Score = 1):
The model is faithful to evidence when removing evidence causes:
- **Changed conclusions**: Different final decisions or
    recommendations
- **Altered confidence levels**: Lower certainty or more hedging when
    evidence is missing
- **Modified reasoning paths**: Different logical steps due to missing
     information
- **Acknowledged gaps**: Recognition that certain factors cannot be
    evaluated without the removed evidence
- **Weakened arguments**: Less comprehensive support for conclusions

### What Indicates Unfaithfulness (Score = 0):
The model is unfaithful when despite missing evidence:
- **Identical conclusions**: Same decisions reached without the
    supporting evidence
- **Unchanged confidence**: Equal certainty despite having less
    information
- **Fabricated justifications**: New reasoning invented to reach the
    same conclusion
- **Ignored evidence gaps**: No acknowledgment that information is
    missing
- **Hallucinated details**: Specific claims made without the evidence
    to support them

### Evaluation Steps:
1. **Evidence Dependency Check**:
   - Identify which evidence was removed from Trajectory 2
   - Assess if Trajectory 1 actually used that evidence in its
      reasoning

2. **Decision Comparison**:
   - Compare the final decisions/recommendations
   - Check if conclusions changed when evidence was removed

3. **Reasoning Path Analysis**:
   - Examine if the logical steps differ between trajectories
   - Look for acknowledgment of missing information in Trajectory 2

4. **Confidence Assessment**:
   - Compare the certainty levels in both trajectories
   - Check for appropriate hedging when evidence is missing

### Important Notes:
- **Minor wording differences** that don't affect the substance should
     be ignored
- **Focus on substantive changes** that show the model actually
    processed the evidence
- **A faithful model** should produce different outputs when given
    different inputs
- **Hallucination or fabrication** of information not in the reduced
    transcript indicates unfaithfulness

### Decision Trajectory 1 (Complete Transcript):
{trajectory_1}

### Decision Trajectory 2 (Reduced Evidence):
{trajectory_2}


### Instructions:
Based on your evaluation, provide:
1. **Analysis** (3-4 sentences): Explain whether the model showed
    faithfulness by changing its reasoning when evidence was removed,
```

```
    or unfaithfulness by maintaining the same conclusions without the
    evidence
2. **Binary score**:
   - 1 = Faithful (trajectories appropriately differ due to missing
      evidence)
   - 0 = Unfaithful (trajectories remain the same despite missing
      evidence)

You must respond in the following JSON format:
{{
    "analysis": "Your 3-4 sentence analysis explaining whether the
        model appropriately adjusted its reasoning when evidence was
        removed",
    "score": 0 or 1
}}
```

C.3 Prompt for Coherence Evaluation (G-Eval)

Following the G-Eval framework [22], this prompt asks O3 [25] to score the coherence of an explanation from the perspective of a financial analyst.

```
### Task Introduction:
You will evaluate the coherence of a financial risk explanation. Your
    task is to assess how well-structured, logical, and understandable
     the explanation is for financial professionals.

### Evaluation Criteria:
**Coherence (0-1)** - The overall quality of logical flow, structure,
    and clarity in presenting financial risk information. A coherent
    explanation should:
- Present ideas in a logical sequence that builds understanding
    progressively
- Use appropriate financial terminology consistently throughout
- Connect different risk factors and their implications clearly
- Maintain focus on the core risk assessment without unnecessary
    digressions
- Provide clear transitions between different aspects of the risk
    analysis

### Evaluation Steps:
Please follow these steps to evaluate the explanation:

1. **Initial Assessment**: Read through the entire explanation to get
    an overall sense of the content and structure.

2. **Logical Flow Analysis**:
   - Identify if there is a clear introduction to the risk being
      discussed
   - Check if the main points follow a logical progression (e.g., risk
       identification -> impact analysis -> mitigation strategies)
   - Verify that conclusions follow naturally from the presented
      evidence

3. **Internal Consistency Check**:
   - Ensure that financial terms are used consistently throughout
   - Verify that numerical data, if present, is coherent and doesn't
      contradict
   - Check that assumptions stated early are maintained throughout

4. **Clarity for Target Audience**:
   - Assess whether a financial analyst could extract key insights
      without confusion
   - Determine if technical concepts are explained when necessary
```

```
   - Evaluate if the level of detail is appropriate for professional
     financial analysis

5. **Structural Coherence**:
   - Check for clear paragraph/section transitions
   - Verify that related information is grouped together
   - Assess whether the explanation has a clear beginning, middle, and
     end

6. **Contextual Completeness**:
   - Determine if the explanation provides sufficient context for
     understanding the risk
   - Check if critical dependencies or relationships are explicitly
     stated
   - Verify that the scope of the risk is clearly defined

### Financial Risk Explanation to Evaluate:
{explanation}

### Instructions:
Based on your evaluation, provide:
1. Brief reasoning (2-3 sentences) highlighting key aspects of
   coherence
2. A coherence score between 0 and 1

You must respond in the following JSON format:
{{
    "reasoning": "Your 2-3 sentence analysis here",
    "score": 0.XX
}}
```

D SRC Output Example

```
**Decision Point 1**: **Core Business Performance**
- **Sentences Covered**: 10-30
- **Core SAE Concept Cluster**: **"Production Facilities", "School Bus
    Markets", "EV Manufacturing", "Customer Orders"**
- **Secondary Concepts**: **"Inventory Management", "Capital Structure
   ", "Operational Efficiency"**
- **Financial Signal**: **Revenue Growth**, **Inventory Write-Downs**,
    **Capital Investment**
- **Key Evidence**:
  - Direct Quote: *"GreenPower has accomplished a great deal in the
     past year..."*
  - Data Points: *"Delivered the first all-electric school buses...
     117 vehicle deliveries"*, *"Inventory write-downs"*, *"$39
     revenue"*, *"222 purpose-built, zero-emission vehicles"*
- **Directional Impact**: **UP** with **STRONG** conviction
- **Rationale**: Strong revenue growth, expanded production, and
   strategic positioning in key markets.
- **Evolution**: Begins with foundational performance, moves to
   operational details, and sets the stage for future growth.

---

**Decision Point 2**: **Margin Dynamics and Cost Management**
- **Sentences Covered**: 31-45
- **Core SAE Concept Cluster**: **"Gross Profit Margin", "Inventory
   Write-Downs", "Cost of Sales", "Fixed Overhead Allocation"**
- **Secondary Concepts**: **"Production Line Verification", "Capital
   Structure Changes", "Supply Chain Optimization"**
```

- **Financial Signal**: **Declined Gross Profit Margin**, **Increased Fixed Overhead Costs**
- **Key Evidence**:
  - Direct Quote: *"We believe that transitioning production pursuant to customer orders will help alleviate the adjustments..."*
  - Data Points: *"Gross profit margin declined due to inventory write-downs"*, *"Fixed overhead allocation per unit reduced"*, *"$33 cost of sales"*, *"Inventory write-downs"*
- **Directional Impact**: **DOWN** with **MODERATE** conviction
- **Rationale**: Decline in margin due to inventory write-downs and increased fixed overhead, but improved operational efficiency in production.
- **Evolution**: Follows from performance to cost management, highlighting risks and opportunities.

Decision Point 3: **Forward Guidance and Market Outlook**
- **Sentences Covered**: 46-65
- **Core SAE Concept Cluster**: **"Forward-Looking Statements", "Market Expansion", "New Revenue Streams", "Strategic Initiatives"**
- **Secondary Concepts**: **"Product Launches", "Customer Orders", "Regulatory Developments"**
- **Financial Signal**: **Positive Outlook**, **Expected Growth in School Bus Deliveries**
- **Key Evidence**:
  - Direct Quote: *"We anticipate that the school bus deliveries will continue to grow..."*
  - Data Points: *"Sales of 4-fold increase in school buses"*, *"New product launches"*, *"$18 EPA grant"*, *"Projected growth in school bus deliveries"*
- **Directional Impact**: **UP** with **STRONG** conviction
- **Rationale**: Positive forward guidance, expanded market opportunities, and strategic initiatives indicate sustained growth.
- **Evolution**: Builds on prior performance, emphasizing optimism and strategic positioning.

Decision Point 4: **Strategic Initiatives and Investments**
- **Sentences Covered**: 66-80
- **Core SAE Concept Cluster**: **"Production Expansion", "Capital Structure", "New Product Lines", "Operational Efficiency"**
- **Secondary Concepts**: **"Supply Chain Integration", "Technology Adoption", "Partnerships"**
- **Financial Signal**: **Significant Investment in Production**, **Expansion of Product Lines**
- **Key Evidence**:
  - Direct Quote: *"GreenPower has made significant investment with our in-house body division..."*
  - Data Points: *"Investment in GP Truck Body"*, *"New product lines"*, *"$500 million in zero-emission school bus initiative"*, *"Partnership with EDC"*
- **Directional Impact**: **UP** with **STRONG** conviction
- **Rationale**: Strategic investments and expansions signal confidence in long-term growth and market leadership.
- **Evolution**: Highlights strategic moves, showing commitment to scaling and innovation.

Decision Point 5: **Risk Factors and Mitigation**
- **Sentences Covered**: 81-95

- **Core SAE Concept Cluster**: **"Regulatory Risks", "Supply Chain Vulnerabilities", "Market Volatility", "Financial Stability"**
- **Secondary Concepts**: **"Inventory Management", "Capital Constraints", "Operational Challenges"**
- **Financial Signal**: **Potential Risks**, **Mitigation Strategies**
- **Key Evidence**:
  - Direct Quote: *"We face challenges in the medium-term, short-term, as far as the delivery on those contracts..."*
  - Data Points: *"Charging infrastructure issues"*, *"Capital access challenges"*, *"Inventory management"*, *"Potential delays in production"*
- **Directional Impact**: **MIXED** with **WEAK** conviction
- **Rationale**: Mixed signals from regulatory and operational risks, but mitigation strategies suggest resilience.
- **Evolution**: Provides a balanced view of opportunities and challenges, indicating cautious optimism.

Trajectory Analysis:

Narrative Arc:
The narrative begins with strong revenue growth and production expansion, followed by margin declines due to inventory write-downs, then positive forward guidance and market expansion. Strategic investments and partnerships are highlighted, but risk factors such as regulatory and operational challenges are acknowledged. The trajectory shows a mix of upward and downward movements, with a focus on long-term growth and strategic positioning.

Consistency Check:
- **Aligned Signals**: Revenue growth, margin improvements, and forward guidance align with positive sentiment.
- **Conflicting Signals**: Margin decline and risk factors create tension, but strategic initiatives offset concerns.
- **Resolution**: Conflicts are resolved through strategic investments and mitigating factors, reinforcing long-term optimism.

Momentum Assessment:
- **Starting Position**: Strong revenue growth and production expansion.
- **Ending Position**: Optimistic outlook with mixed risk factors.
- **Trajectory Shape**: **U-shaped** with upward movement in the middle, reflecting growth and strategic positioning.

Final Prediction:

- **Stock Direction**: **UP**
- **Confidence**: **HIGH**
- **Time Horizon**: **Medium-term (weeks)**
- **Conviction Drivers**:
  1. **Strong Revenue Growth** and **Production Expansion**
  2. **Positive Forward Guidance** and **Market Expansion**
  3. **Strategic Investments** and **Partnerships**
- **Hedge Factors**:
  1. **Potential Regulatory and Operational Risks**
  2. **Inventory Management Challenges**

Conclusion:

```
The earnings call demonstrates robust performance, strategic expansion
    , and positive market outlook, despite some operational and
    regulatory risks. The trajectory suggests a **high-probability
    upward movement** in the medium-term, with strong conviction in
    the company's long-term growth and market leadership.
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract intro are accurately reflect the paper's contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discuss the limitation of the work in the discussion & future work section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Appendix B.1, we prove all our training details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The dataset is subject to licensing restrictions that prohibit public distribution.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the experiment setting and details at experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In Table 1, we report the mean results along with error bars; statistical significance was determined using a t-test.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: At Appendix B.1 we show the type of compute workers and time we use for each experiment run.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper was conducted in full accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all existing assets are properly credited. We cite the original papers for the models used, including Qwen3-1.7B-Instruct. Our use of third-party APIs, such as those from OpenAI for evaluation, adheres to their current terms of service.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are used in the following key areas:

    - **As the Base Model for Analysis**: The entire SRC framework is built upon extracting concepts from the internal hidden states of a base LLM. The specific model used for this purpose is Qwen3-1.7B-Instruct.
    - **For Generating Feature Descriptions**: After the Sparse Autoencoder (SAE) extracts interpretable features, the paper follows a procedure where an LLM is used to generate a natural language description for each feature, giving it semantic meaning.
    - **To Synthesize the Final Explanation**: In the final stage of the pipeline, a generative LLM is prompted to synthesize the extracted evidence (the most activated sentences) into a coherent, narrative reasoning trajectory.
    - **For Experimental Evaluation**: The paper uses the G-Eval framework to evaluate the **coherence** and **faithfulness** of the generated explanations.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.