LLM for Zero-Shot Diachronic Semantic Shift Detection

Anonymous ACL submission

Abstract

This paper explores the use of hidden states from large language models (LLMs) to detect semantic shifts in specialized domains via a zero-shot approach. While encoder-based models dominate this research, they face limitations in context length, computational cost, and interpretability. We propose extracting contextualized word embeddings from the decoder hidden states of Llama 3 series models (Dubey et al., 2024). Our method employs structured input formulations to guide LLMs in generating context-sensitive word definitions, from which we extract hidden state representations. Using a historical corpus (Credit Suisse Bulletin, 1970–2018), we measure semantic shifts with Jensen-Shannon divergence. Experimental results show decoder hidden states effectively capture contextualized semantics, demonstrated by a case study of the word "interest". To our knowledge, this is the first study leveraging decoder hidden states prompted by definition generation without reliance on generated text analysis. Our method enables decoder-only models to effectively detect semantic shifts, providing a computationally efficient, interpretable alternative for unlabeled data while significantly reducing computational overhead compared to encoder-based approaches.

1 Introduction

In recent years, Transformer-based language models have advanced significantly in modeling representation through dynamic embeddings, enabling increasingly sophisticated analyses in downstream tasks such as semantic shift detection (SSD). SSD is the task of identifying changes in word meanings between two sets of texts, such as diachronic or text-genre corpora, which is crucial for understanding how language evolves. The evolution of SSD methods has progressed from static word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) to dynamic representation models. While encoder-based models now dominate current SSD research (Montanelli and Periti, 2023; Rudolph and Blei, 2017; Ishihara et al., 2022), their limitations excite exploration of alternative model architectures. Diachronic domain-specific corpora are valuable, but the expensive manual annotation impedes their linguistic research.

Existing SSD research predominantly relies on encoder-based models, which have three main constraints for SSD. First, commonly used encoder-only models have modest parameters and pretraining volumes compared to LLMs (e.g., BERT-Large 330M vs. DeepSeek V3 671B (Devlin et al., 2019; Liu et al., 2024)), suggesting a disparity in representational ability. Second, applying BERT-like models often requires transfer learning pipelines and side-tricks (Gao et al., 2021; Ishihara et al., 2022). Recent work using decoder-only models could address its unidirectional nature by loading unmasked LLMs (e.g., NV-Embed (Lee et al., 2024), LLM2Vec (BehnamGhader et al., 2024)); however, this approach blurs the border between decoder and encoder models. Other studies focus on analyzing generated contents (Giulianelli et al., 2023; de Sá et al., 2024), which lack scalability. Therefore, an efficient SSD solution that can leverage LLMs for large-scale unlabelled corpora remains challenging.

This paper focuses on leveraging LLMs for zero-shot SSD in specialized domains, addressing a crucial question: how can we extract semantic information without fine-tuning overhead? Our zero-shot approach eliminates dependency on gold-standard datasets, making it applicable to any historical texts. We hypothesize that decoder-only LLMs' hidden states contain sufficient semantic information to detect meaning shifts without modification or post-training steps.

To address the above challenges, we propose an innovative approach that utilizes the LLM (LLaMA 3 series (Dubey et al., 2024)) for zero-shot SSD

by extracting decoder hidden states directly from the model. Our method employs structured input formulations to guide LLMs in generating context-sensitive word definitions, from which we extract semantics. This approach leverages the LLMs' intensive pre-training and extended context capabilities without specialized post-training. To our knowledge, this is the first study leveraging decoder hidden states guided by prompted definitions, distinguishing our approach clearly from prior decoder-based research. We analyze the evolution of semantics over 50 years (1970–2018) using a historical corpus (Credit Suisse Bulletin) (Volk et al., 2016). Experimental results show that decoder hidden states effectively capture contextualized semantics, demonstrated by a case study.

Our contributions are:

- A novel zero-shot SSD method utilizing decoder hidden states instructed by definition generation, without requiring fine-tuning or analysis of generated content.
- A computationally efficient, interpretable alternative to encoder-based methods for semantic shift analysis in unlabeled corpora.
- Validation of the method's effectiveness through analysis of semantic evolution in a 50-year financial corpus.

The rest of this paper is organized as follows: Section 2 summarizes related work on semantic shift detection and LLM applications. Section 3 describes our proposed method for extracting and analyzing hidden states. Section 4 presents our experimental setup and results, followed by discussions of implications in Section 5. Finally, Section 6 concludes the paper and suggests directions for future work. The code for this work is available on GitHub (anonymized for review).

2 Related Work

Semantic Shift Detection (SSD) has evolved from static word embeddings to context-aware models. Encoder-based models like BERT (Devlin et al., 2019) have dominated recent SSD research (Montanelli and Periti, 2023), with various techniques for improving performance (Giulianelli et al., 2020; Ishihara et al., 2022). These approaches benefit from bidirectional context but face significant limitations: they typically have smaller base models, require resource-intensive fine-tuning for diachronic adaptation, and struggle with limited contexts, limiting their effectiveness in SSD.

Decoder-only LLMs remain largely unexplored for SSD despite their compelling advantages: massive parameter scales and strong zero-shot capabilities across domains (Brown et al., 2020; Dubey et al., 2024). Limited work with decoder models primarily focuses on analyzing generated text (Wang and Choi, 2023; de Sá et al., 2024) rather than utilizing their internal representations. Some approaches like model modification (Lee et al., 2024; BehnamGhader et al., 2024) have been proposed to adapt LLMs for embedding extraction, but these modifications blur the distinction between decoder and encoder architectures, adding unnecessary complexity.

Our work addresses a significant research gap by pioneering the direct utilization of decoder hidden states for zero-shot SSD—a novel approach in the literature. Unlike prior work that analyzes generated outputs, we propose a more elegant solution: employing structured prompts to extract semantically rich representations from decoder hidden states without generation analysis or model modification. This approach uniquely combines the representational power of LLMs with computational efficiency, offering the first truly zero-shot SSD method that requires neither fine-tuning nor gold-standard datasets while maintaining interpretability.

3 Methodology

3.1 Overview

This section details the method for extracting word embeddings by leveraging the hidden states of a decoder-only LLM architecture for zero-shot Semantic Shift Detection (SSD). Since decoders are generally considered to only encode unidirectional information from left to current token, their hidden states are typically not considered suitable for contextualized word embedding (CWE) extraction. Our proposed method uses the hidden states of LLMs to represent CWEs by guiding the model with carefully designed prompts, avoiding specialized post-training.

3.2 Model and Embedding Extraction

3.2.1 Embedding Extraction Techniques

We investigated five strategic positions for extracting contextualized word embeddings (CWEs) from the model's hidden states:

- **input_mean**: Average of all input sequence hidden states.
- **input_last_token**: Hidden state of the final input token (conditioning the first generated token).
- **eos_token**: Hidden state of a manually added EOS token at the input end.
- **output_mean**: Average of generated definition hidden states.
- **output_eos**: Hidden state of the model-generated EOS token.

We extracted hidden states from all layers and compared their semantic representation capabilities. Based on comparative evaluation of polysemous word clustering clarity (Section 4.3), `input_last_token` outperforms other positions in distinguishing polysemous meanings, while requiring no additional model modifications and enabling extraction without generation steps. This position was selected for subsequent analyses.

3.3 Structured Input Formulation for Definition Elicitation

We systematically formulated structured in-context learning templates to elicit context-sensitive word definitions from the LLM. Through comparative evaluation (Section 4.4), we identified a three-role dialog template (`icl_basic`), consisting of a system role defining the task, a user role providing an example and query, and an assistant role for generation, that maximized definition consistency and minimized generation errors. The complete template specifications and error analysis metrics are detailed in Appendices B and C.

3.4 Semantic Shift Detection Techniques

We measure semantic shifts quantitatively using Jensen-Shannon divergence (JSD), a symmetric, bounded metric (0–1) suitable for capturing distributional changes. Unlike cosine similarity, which compares only average vectors, JSD fully leverages our large-scale collection of CWEs, enabled by our zero-shot approach, and evaluates entire embedding distributions, providing enhanced sensitivity to subtle semantic shifts.

3.5 Dimensionality Reduction and Visualization

For visualization and computational efficiency, we use UMAP (McInnes et al., 2018) to reduce the high-dimensional embeddings. Specific UMAP parameters are detailed in Appendix A.3. UMAP was selected for its ability to preserve topological structure and improve cluster separation compared to linear methods like PCA. Our analysis employs scatter plots to identify polysemy and temporal visualizations to track shifts.

4 Experiments and Results

This section details our experimental methodology and presents the findings. The core analyses were performed using the LLaMa 3.1 8B Instruct model. Comprehensive details regarding the experimental setup, model configurations, hyperparameters, and computational resources are provided in Appendix A.

4.1 Datasets

We use a long-span professional publication corpus, the English portion of the *Credit Suisse Bulletin* corpus (Volk et al., 2016) (1970-2018, OCR and PDF sources), which is available for research purposes. The corpus underwent common NLP preprocessing (POS tagging, lemmatization) from raw XML data. This preprocessing was followed by extensive rule-based filtering to enhance data quality and appropriateness. These filters included checks for proper ending punctuation, the presence of emails or URLs, excessive consecutive digits, sentence length constraints (typically 5 to 100 words), non-English character limits, a minimum number of alphabetic characters, and an English language check. The filtering for emails, URLs, and consecutive digits also served to reduce the presence of directly identifiable information in the dataset. Additionally, perplexity filtering was employed to remove nonsensical sentences (see Appendix A.1 for model details). The LLaMA 3 series models are used under their open-source license.

For evaluation purposes, target words in the cleaned sentences were annotated with WordNet (Miller, 1995) sense IDs using the OpenAI O1-mini API via batch processing. WordNet come with expert-curated lexical database that organizes English words into sets of cognitive synonyms (synsets), each representing a distinct lexical meaning structure. We use WordNet to provide sense definitions, enabling structured comparison across time. These annotations serve as a silver standard—machine-assigned labels grounded in expert definitions, offering interpretability but

not gold-standard precision. The final dataset was grouped into 5-year intervals for long-term SSD.

4.2 Optimal Layers for Semantic Representation

We visualized each layer's representational capability for "capital" through UMAP dimensionality reduction. Figure 1 shows scatter plot distributions in four representative layers and WordNet labels.

As the layers deepen, the model gradually forms semantic clusters, showing significant grouping trends from layer 22 onward. Notably, the blue points (representing "wealth in the form of money or property") and purple points (representing "assets available for producing other assets") appear partially overlapped in later layers, reflecting their close semantic relationship. From WordNet's perspective, the blue points represent an extension of the purple points in terms of ownership - while purple points emphasize the productive function of assets, blue points highlight the possession aspect of wealth. This overlap in the LLM's representation space demonstrates how the model captures subtle semantic relationships that align with lexicographical knowledge structures.

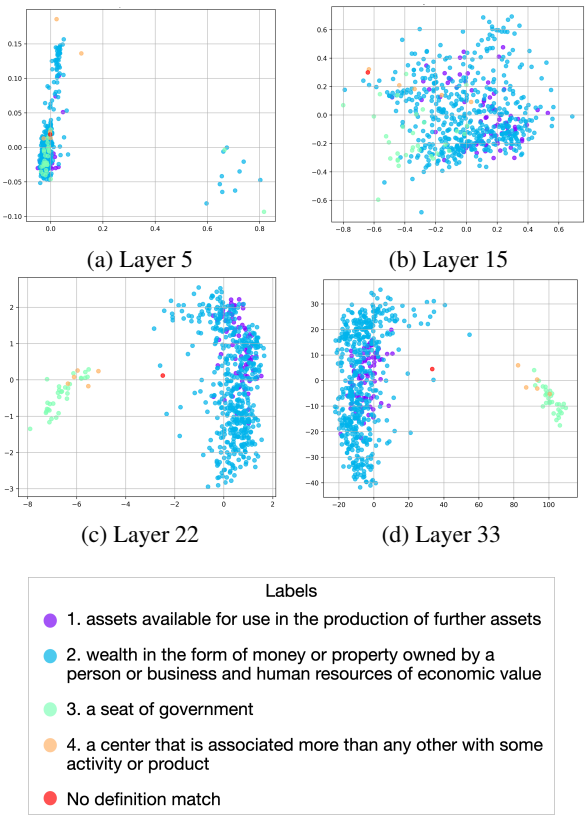
Our analysis of semantic representation across different layers showed that the early layers mainly capture syntactic features, while middle layers begin to cook semantically relevant clusters. The later layersdemonstrate the ability to distinguish different semantic categories, forming distinct cluster structures. Based on this evidence, we concatenated the model's later layers as the primary basis for semantic shift analysis.

4.3 Optimal Embedding Extraction Positions

We visualized UMAP-reduced embeddings for polysemous words to compare extraction methods (`input_mean`, `input_last_token`, `eos_token`, `output_mean`, `output_eos`). Figure 2 shows this comparison using ïrate, utilizing the previously determined optimal later layers (21-33).

The `input_last_token` position shows the clearest separation of semantic clusters. `eos_token` performs reasonably well. `input_mean` tends to blur distinctions. Output-based methods (`output_mean`, `output_eos`) show less coherent clustering. The distinct cluster identified by `input_last_token` corresponds to idiomatic uses (e.g., ät any rate).

We also computed pairwise cosine similarity between the methods. Figure 3 shows the similarity



(a) Layer 5 (b) Layer 15

(c) Layer 22 (d) Layer 33

(e) WordNet Labels

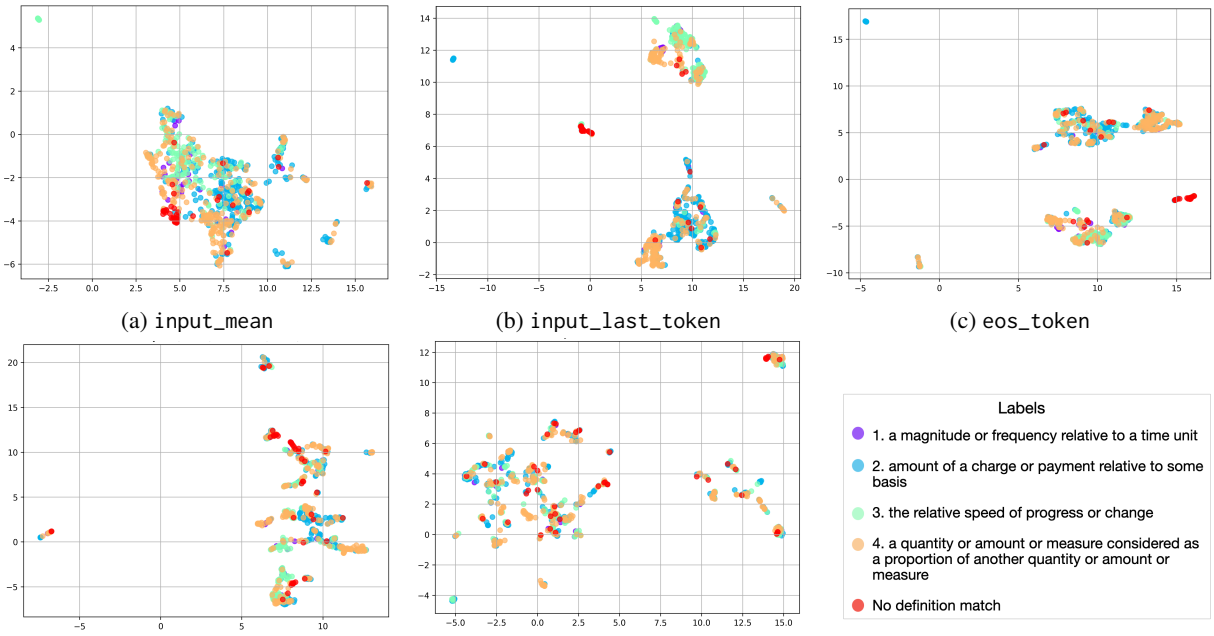
Figure 1: Semantic representation capability for "capital" at representative layers (5, 15, 22, 33) and WordNet label distribution (Llama 3.1 8B), colored by WordNet sense ID.

matrix.

Based on the clarity of semantic clustering and distinctiveness from other methods, `input_last_token` is selected as the optimal extraction position for subsequent analyses.

4.4 Evaluation of Instruction Design

We compared two prompt designs: `icl_basic` (Table 2) and `icl_context_aware`. Prompts were evaluated based on their tendency to produce erroneous outputs, specifically examining common error categories including Content duplication (E1 CD), Polysemic enumeration (E2 PE), Instruction Response Shift (E3 IRS), and CoT activation (E4 CoTA). Table 1 summarizes the performance. The `icl_basic` prompt demonstrated more stable performance, minimizing most errors (E1, E3, E4) while generating concise definitions, despite occasional polysemic enumeration (E2), and was therefore chosen for the main experiments. A summarized description for each error type is provided in Appendix C.



(a) `input_mean` (b) `input_last_token` (c) `eos_token`

(d) `output_mean` (e) `output_eos` (f) Rate labels

Figure 2: Comparison of semantic clustering capabilities of five different embedding extraction methods for the word ŕateüsing stacked later layers (21-33). Colours correspond to WordNet definitions. Red points are rejected for labelling by O1-mini.

Prompt	E1: CD	E2: PE	E3: IRS	E4: CoTA
icl_basic	✗	✓	✓	✗
icl_context_aware	✗	✓	✓	✓

Table 1: Observed error behaviours for compared prompts. ✓indicates the error was observed, ✗ indicates it was not.

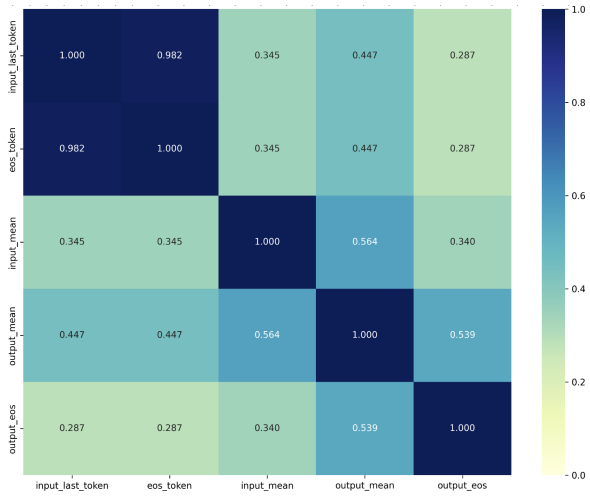


Figure 3: Pairwise cosine similarity heatmap between different embedding extraction methods. Darker color indicates high similarity.

4.5 Model Configuration Analysis

Results confirmed that decoder hidden states from later layers (21-33) of Llama 3.1 8B effectively capture context-specific semantics. Our ablations revealed: (1) the 8B model outperformed the 3B model, demonstrating clear scaling benefits; (2) sufficient context length (5 sentences) improved differentiation of subtle meanings despite higher computational costs; and (3) the `icl_basic` prompt balanced guidance and stability better than more complex alternatives. These findings suggest that larger models with adequate context window benefit semantic representation quality. Detailed ablation results are provided in Appendix F.

4.6 Semantic Shift Detection

4.6.1 Overall Semantic Shift Trajectory Visualization

To visualize the semantic evolution of financial terminology over five decades, we mapped the trajectories of 44 financial terms in a shared semantic space using UMAP dimensionality reduction on hidden states extracted via the `input_last_token` method. Figure 4 presents these trajectories, where each point represents a term's semantic centroid within a specific time window (1970-2018).

The visualization reveals varying trajectory

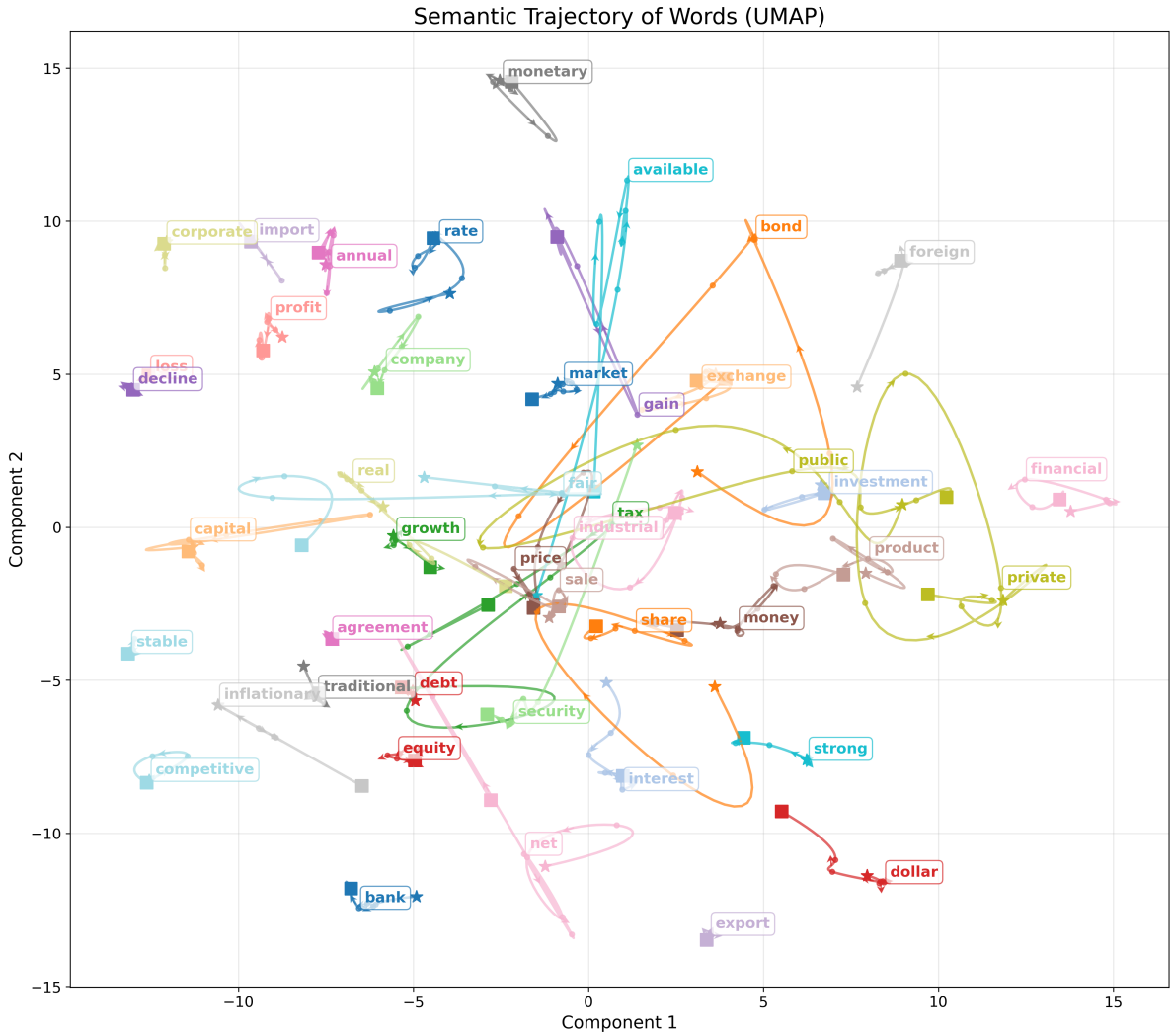


Figure 4: Semantic shift trajectory map of financial terms (1970-2018). Each trajectory represents a word's semantic shift path across time windows.

lengths and patterns among the terms. Some terms exhibit short, concentrated trajectories (e.g., "bank", "profit"), while others show longer, more directional movements (e.g., "security", "bond") or fluctuating patterns without clear directional trends (e.g., "monetary"). The quantitative analysis of these patterns is presented in Section 5.

4.6.2 Quantitative Analysis of "Interest"

Figure 5 presents the sense distribution of "interest" across time periods (1970-2018). The data shows two dominant meanings: definition 4 ("fixed charge for a service, usually a certain percentage of the loan amount") and definition 1 ("attention and curiosity towards someone or something").

The data shows that between 1970 and 1995, definition 4 represented over 75% of usage instances. After 2004, definition 1 increased from under 20% to nearly 50%. The line graph shows the overall frequency of the term declined during this period.

Figure 6 presents the Jensen-Shannon divergence (JSD) measurements for "interest." The JSD values remained at approximately 0.15 from 1970 to 1990, then increased to 0.4 by 2008.

Additional quantitative results for the term "real" are provided in condensed form in Appendix E.

5 Discussion

5.1 Interpretation of Semantic Trajectories

Based on the trajectory visualization in Section 4.6.1, we identified three distinct patterns of semantic evolution:

- **Stability**: Terms such as "bank" and "profit" maintain consistent semantic positions, exhibiting short, concentrated trajectories.
- **Shifting**: Terms such as "security" and "bond"

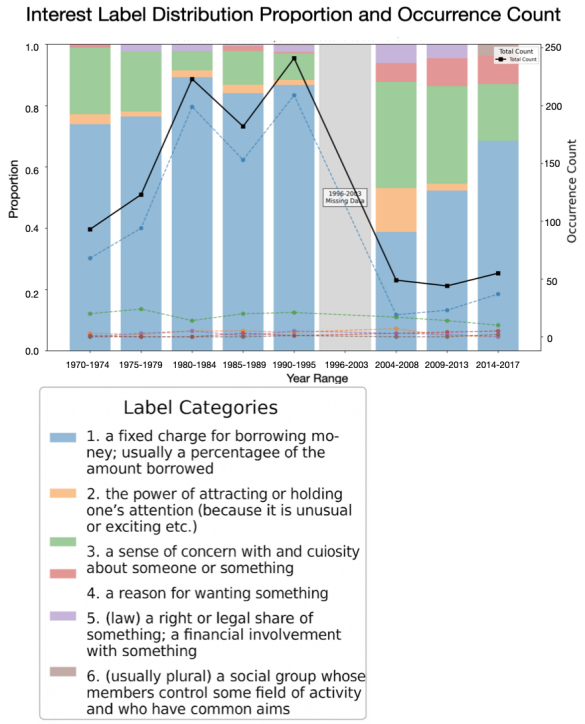


Figure 5: Sense label distribution of "Interest" (1970-2018). The bar chart shows the proportion of each sense in different periods, and the line chart represents the word frequency changes.

show unidirectional semantic shift with clear directional movement.

- **Oscillation**: Terms such as "monetary" display fluctuating patterns without a clear directional trend, reflecting periodic adjustments or sampling variability.

This trajectory analysis provides valuable insights into the semantic relationships between financial terms and enables comparative analysis across multiple terms. The map reveals both gradual shifts over decades and more pronounced transitions in specific periods.

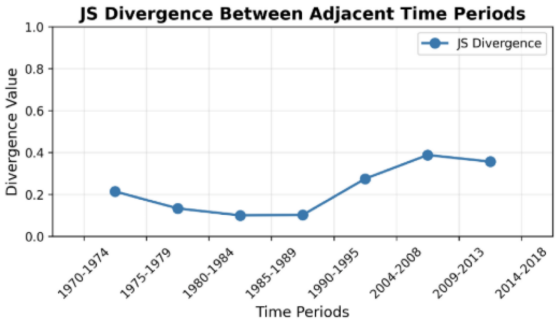


Figure 6: JS divergence for "interest" (1970-2018). The figure shows JS Divergence over time, indicating semantic shift intensity.

5.2 Analysis of "Interest" Semantic Evolution

The case study of "interest" in Section 4.6.2 demonstrates a significant semantic transition from specialized financial meaning toward more general usage. The sharp increase in JSD values after 1990 corresponds precisely to the observed shift from financial to general meaning. This indicates that our method effectively captures meaningful semantic evolution rather than random variation. The quantitative data reveals a clear trend: the financial sense (definition 4) clearly dominated initially, representing over 75% of usage instances, while after 2004, a significant shift occurred—the general sense (definition 1) increased dramatically, approaching equivalence with the financial meaning.

The divergent semantic trajectory of "interest" likely reflects changing content focus in the Credit Suisse Bulletin over this period, potentially indicating a shift in communication strategy from purely financial reporting toward broader topics with general audience appeal. This exemplifies a broader "universalization" trend observed across multiple specialized terms in our corpus.

5.3 WordNet Labels vs. LLM Representations

Comparing WordNet semantic annotation with LLM-derived clustering reveals complex relationships. WordNet provides structured reference, but its distinctions do not always align with LLMs' semantic space. Key patterns observed: (1) LLMs exhibit broader semantic clustering than WordNet's fine distinctions; (2) LLMs sometimes differentiate instances with identical WordNet senses based on context; and (3) LLMs distinguish idiomatic expressions not specifically in WordNet. These patterns suggest LLMs capture broader category groupings and subtle contextual distinctions reflecting actual language use, which may differ from WordNet's predefined classifications.

5.4 Broader Trends in Domain Terminology

Our analysis across multiple terms reveals a prevalent pattern in specialized domains: the gradual transition of terminology from restricted technical usage toward more generalized applications over time. This "universalization" phenomenon extends beyond individual cases like "interest" (Section 4.6.2) and "real" (Appendix E).

We attribute this pattern to several potential factors:

- Changes in publication strategy, with increased focus on broader audiences
- Evolution of financial discourse toward more accessible language
- Semantic broadening as specialized terms enter mainstream usage

This observed trend highlights how technical language naturally evolves to serve both specialized and general communication functions over time, particularly in long-running professional publications like the Credit Suisse Bulletin.

5.5 Methodological Implications for SSD

Our approach provides robust, interpretable embeddings from decoder hidden states without fine-tuning, making it particularly suitable for large-scale historical corpus analysis. The JSD metric effectively quantifies semantic shifts, demonstrating practical utility for identifying both subtle and dramatic meaning changes over time.

The case studies collectively validate our analytical framework's capability to capture different change patterns, including semantic stability, shift, and differentiation. The combination of hidden state extraction and distributional analysis offers a powerful, computationally efficient method for semantic evolution research in unannotated diachronic corpora, with potential applications beyond the financial domain.

5.6 Limitations

This study has limitations. Regarding data, the diachronic corpus faces challenges: the inherent sparsity of language combined with this leads to infrequent appearance of some interesting terms, resulting in insufficient sample sizes for stable statistical metrics and semantic traces. As a weak label source, WordNet suffers from subjectivity and granularity issues that differ from the semantic structures revealed by LLMs. Methodological limitations are constituted by the quality of model generation, the selection of hidden state layers, dimension loss, and the lack of strong labels for precise evaluation. Additionally, the generated text still contains errors such as content repetition and multiple-definition listing, which may affect the quality of contextualized word embedding.

6 Conclusion and Future Work

6.1 Summary of Contributions

This paper presents a novel zero-shot method for SSD using decoder hidden states from LLMs (Llama 3), guided by definition generation using structured inputs. We demonstrated its effectiveness in capturing semantic shifts in a 50-year diachronic corpus without fine-tuning. Key contributions include: (1) pioneering the combined use of decoder hidden states and definition generation guided by structured inputs for embedding-based SSD, (2) proposing an input formulation technique for embedding extraction, (3) applying JSD for distributional shift analysis, and (4) providing insights into long-term semantic evolution. Our approach offers a computationally efficient alternative to traditional methods and serves as a methodological advancement for SSD in any unlabeled corpus. Our approach also demonstrates how leveraging large decoder-only language models can shift methodological paradigms within computational linguistics research. Our results highlight the complementarity between decoder-only LLM representations and structured semantic resources like WordNet, offering broad potential applications beyond financial semantic analysis as a versatile analytical tool in linguistics, artificial intelligence, and data science.

6.2 Future Works

Future work includes: (1) Optimizing prompt design and potentially using generated definitions as pseudo-labels, freeing us from our dependence on WordNet's auxiliary. (2) Investigating reasoning models with explicit think tokens, such as DeepSeek R1, to extract CWEs after their intermediate reasoning steps, potentially leading to more accurate and interpretable embeddings. (3) Extending to cross-lingual semantic shift detection by leveraging the aligned, parallel translations available in our dataset, enabling both language-specific SSD and comparative analysis of semantic evolution across languages. (4) Improving computational efficiency by leveraging KV caching to enable parallel extraction of hidden states for all words in one sentence. Combining LLM representations with structured knowledge like WordNet may also yield benefits.

References

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Semantic change characterization with llms using rhetorics. *arXiv preprint arXiv:2407.16624*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. *arXiv preprint arXiv:2004.14118*.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. *Preprint*, arXiv:2305.11993.

Shotaro Ishihara, Hiromu Takahashi, and Hono Shirai. 2022. Semantic shift stability: Efficient way to detect performance degradation of word embeddings and pre-trained language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 205–216, Online only. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

George A Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection. *arXiv preprint arXiv:2304.01666*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Maja Rudolph and David Blei. 2017. Dynamic bernoulli embeddings for language evolution. *arXiv preprint arXiv:1703.08052*.

Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016. Building a parallel corpus on the world's oldest banking magazine.

Ruiyu Wang and Matthew Choi. 2023. Large language models on lexical semantic change detection: An evaluation. *arXiv preprint arXiv:2312.06002*.

A Computational Details

A.1 Model Size and Computational Budget

The primary language models employed in this research were Meta's Llama 3.1 8B (approximately 8 billion parameters) and Llama 3.2 3B (approximately 3 billion parameters), accessed via the Hugging Face 'transformers' library. All experiments were conducted on a SLURM system equipped with 4 Nvidia GH200 GPUs. The Llama 3.1 8B model required approximately 6 hours of computation time for the full runs, while the experimental runs for the Llama 3.2 3B model required approximately 3.5 GPU hours, with both utilizing the four GPUs in parallel.

A.2 Experimental Setup and Hyperparameters

For the generation of contextualized definitions and subsequent hidden state extraction, a fixed experimental configuration was utilized. Operations were

performed using ‘bf16‘ mixed precision. Text generation employed a beam search strategy with 3 beams, a temperature setting of 0.3 to ensure deterministic and focused outputs, and a maximum limit of 50 newly generated tokens per definition. No systematic hyperparameter search was conducted; these settings were established based on qualitative assessments from preliminary experiments.

A.3 Data Preprocessing and Software Parameters

Initial data preprocessing included an English language filtering step using the ‘langdetect‘ library with an English probability threshold of 0.7. For dimensionality reduction and visualization, the ‘umap-learn‘ Python package was used. UMAP was configured with ‘n_components=2‘ for 2D visualizations and ‘n_components=16‘ for dimensionality reduction prior to JSD computation; other UMAP parameters (e.g., ‘n_neighbors‘, ‘min_dist‘, ‘metric‘) were kept at their default values from the ‘umap-learn‘ package. Linguistic preprocessing (Part-of-Speech tagging and lemmatization) relied on the spaCy library. WordNet sense annotation was performed using the OpenAI O1-mini model. Visualizations were generated using Matplotlib.

B Prompt Templates

Table 2 shows the primary in-context learning prompt template (`icl_basic`) used for definition generation. We also experimented with a context-aware variant (`icl_context_aware`), the detailed content of which is available on GitHub.

C Response Error Comparison

This appendix summarizes the error types mentioned in Section 4.4. Table 3 provides a brief description and a condensed example for each error type.

D Illustrative Prompt Output Comparison

This section provides illustrative examples of outputs generated by the `icl_basic` and `icl_context_aware` prompt templates (Appendix B). These examples highlight typical qualitative differences and associated error behaviours summarized in Appendix C.

E Additional Case Studies

E.1 Semantic Shift Analysis of "Real"

The term "real" shifted from a specific financial meaning (inflation-adjusted) to a more general usage after 1995. This shift, illustrated by JSD trends (Figure 7), reflects the corpus’s content evolution. JSD analysis showed high values initially, stabilizing after the 1990s as the term’s usage evolved from predominantly technical economic contexts to more general senses.

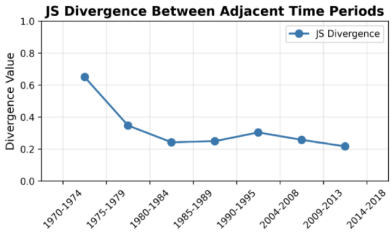


Figure 7: JSD trend for the word "real" across time periods.

F Model Configuration Analysis Details

This appendix provides condensed additional details for the model configuration analysis mentioned in Section 4.5.

F.1 Model Size Comparison: Llama 3.2 3B vs Llama 3.1 8B

In model size comparisons, we observed that Llama 3.1 8B generally produced more accurate and contextually relevant definitions compared to the smaller Llama 3.2 3B model. The 8B model demonstrated better handling of complex contexts and polysemy, though it occasionally added redundant details (Error E3: IRS). The 3B model, while faster, more frequently exhibited content duplication (Error E1: CD) and produced less nuanced definitions.

F.2 Sentence Level Context Length

Longer context (e.g., 5 sentences) consistently improved definition precision for terms like "market" compared to shorter or no context scenarios when using Llama 3.1 8B. Sufficient context helped the model capture more nuanced, systemic professional meanings rather than basic or overly general interpretations. When examining the impact on the term "market" with Llama 3.1 8B, we observed that with no context, the model provided basic

Role	Content
System	You are a dictionary. Your only task is to provide a single definition by its context within 20 words. Do not include bullet points, steps, explanations, or any text besides the definition itself.
User	Example: Sentence: "The company's revenue showed significant growth in Q4." Word: "growth" An increase in size, quantity, or importance over time. For the sentence: "{sentence}" Define the word "{word}":
Assistant	[empty response to be generated]

Table 2: Basic in-context learning prompt template (`icl_basic`).

Error Type	Short Description	Example (Shortened)
E1: CD	Content duplicated from input.	Input: "He implemented..." Output: "He implemented..."
E2: PE	Lists multiple possible meanings.	Input: "...emerging market." Output: "An area where goods... or... A place...Or a particular industry..."
E3: IRS	Adds extraneous comments.	Input: "...political risk..." Output: "An economic system... (Note: I've kept...)"
E4: CoTA	Begins unnecessary reasoning.	Input: "Not fair..." Output: "I understand... Here are the definitions: 1. For sentence..."

Table 3: Summary of response error types.

Table 4: Illustrative comparison of outputs from `icl_basic` and `icl_context_aware`. Error codes (e.g., E4) refer to Appendix C.

Input (Sentence; Word)	Output from `icl_basic`	Output from `icl_context_aware`
Sentence: "The research paper discussed the rate of inflation." Word: "rate"	"A measure, quantity, or frequency, typically compared against another quantity." (Concise, general definition)	"The speed at which economic indicators, such as inflation, change." (Context-specific, exhibit E4: CoTA)
Sentence: "He decided to bank the money he won." Word: "bank"	"To deposit money in a financial institution." (Direct, concise definition)	"1. To deposit money. 2. To rely upon. Here, meaning 1 applies." (Exhibit E2: PE)
Sentence: "The company's capital was invested wisely." Word: "capital"	"Wealth or assets owned by a person or organization." (Focused definition)	"Financial resources for business use, distinct from other meanings such as a capital city." (Context-aware, exhibit E3: IRS)

definitions (e.g., "place or situation"); with 1 sentence, it captured limited aspects (e.g., "fluctuating supply/demand"); with 3 sentences, it performed similar to minimal context; and with 5 sentences, it produced more accurate, systemic definitions (e.g., "system of buyers/sellers, price by supply/demand"). While inference speed was not significantly affected by context length, memory usage increased with longer contexts due to expanded KV caches.